UNIVERSITY
OF TASMANIA

HYPERSPECTRAL CLASSIFICATION OPTIMIZATION USING MULTIOBJECTIVE EVOLUTIONARY ALGORITHM

CONY SUMORO

B.Eng. (Mechatronics Engineering)

Submitted in fulfilment of the requirements for the degree of

Master of Philosophy

AUSTRALIAN MARITIME COLLEGE

UNIVERSITY OF TASMANIA

JUNE 2018

DECLARATION OF ORIGINALITY

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyrights.

Cony Sumoro

STATEMENT OF AUTHORITY OF ACCESS

This thesis may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

Cony Sumoro

ACKNOWLEDGEMENTS

"HE has made everything beautiful in its time" – Ecclesiastes 3:11

This thesis would not have been possible without the support of my supervisory team: Hung Nguyen, Fauzi Hardjanto, and Alex Forrest. I would like to express my sincere thanks to them for their support and guidance throughout my entire studies.

I could not have completed this thesis without the help from Dr. Christiaan Roelfsema from University of Queensland and Dr. Hannelie Botha from CSIRO who provide the seagrass spectral data for research.

Special thanks to the Indonesian community in Launceston especially to Fera who cured my homesickness with Indonesian food and to Ayu for making jokes to keep me sane. This endeavour would not be possible without the enormous support from Martin who has been the light in my life which keeps me motivated when I want to give up. I would like to thank Martin for bringing balance in my postgraduate studies.

I am very grateful for my parents, Kenny and Irene ,who instil the desire to pursue higher studies. I am very grateful to them for inspiring me to wander around the world in the pursuit of knowledge. Last but not least, I would like to thank my brother, William and my sister, Ariel for supporting me unconditionally.

ABSTRACT

The development of hyperspectral imagers provides an avenue to exploit it as a bio-optical taxonomic identification tool. In order to classify the species based on their spectral characteristic, spectral similarity measures are usually applied. Five spectral similarity measures are examined: Spectral Correlation Angle (SCA) for calculating the angle of spectral cross-correlation vectors, Spectral Angle Measure (SAM) for the angle between two spectral vectors, Spectral Information Divergence (SID)for measuring the difference in the probability distribution of two spectra, and hybrid measures: SID-SAM and Normalized Spectral Similarity Measure (nSSM). However, currently there is no specific threshold value for each spectral similarity measure that defines positive classification of one specific species. The conventional method by applying either a fixed or an adaptive threshold value is found to be unreliable. The focus of this thesis is to explore the characteristics of different spectral similarity measures and to utilize the MOEA to find the best value for the threshold value. In addition, the research also put forward the parameter test to find the optimum parameter. A machine learning algorithm, SVM is used to compare the performance of MOEA.

Non-Dominated Sorting Genetic Algorithm (NSGA-II) is a variant of MOEA that is recognized as robust variant, and is therefore selected for this research. In the comparative study of 6 type of seagrass and 4 terrestrial plants, the performance of the discriminating threshold is found to be statistically superior to those from adaptive threshold method. The parameters of NSGA-II were quantified and found that crossover, mutation rate, and different initial chromosome seeds to be significant. The regression models were obtained using this information. The best parameters' values were then used to optimize NSGA-II. The optimized NSGA-II performs on-par with the SVM. The obtained results combined with Probability of Spectral Discrimination (PSD) and Power of Spectral Discrimination (PWSD) were used to suggest preferred similarity measures for specific target seagrass.

Table of Contents

Table of Figures

Table of Tables

Nomenclatures

r_i and s_i	Spectral vectors
$D(r_i \| s_i)$	Relative entropy of s_i relative to r_i
$D(s_i \| r_i)$	Relative entropy of r_i relative to s_i
$I_l(r_{il})$	Self-information measure for r_i
$I_l(s_{il})$	Self-information measure for s_i
p_n	Probability vector p
q_n	Probability vector q
L_U	Upwelling radiance
E_d	Downwelling irradiance
P_{FP}	False positive rate
FP	Number of false positive
N_p	Number of positive sample
P_{FN}	False negative rate
FN	Number of false negative
N_N	Number of negative sample
$P_{total\,FN}$	Total probability of false negative
$P_{total\,FP}$	Total probability of false positive
$\prec_n$	Crowded comparison
i_{rank}	Nondomination rank
$i_{distance}$	Crowding distance
F_i	Front i
$c_{i,k}$	i^{th} child with k^{th} component (crossover)
$p_{i,k}$	i^{th} parent with k^{th} component (crossover)
β_k	Random number
η_c	Crossover distribution index
c_k	Child with k^{th} component (mutation)
p_k	Parent with k^{th} component (mutation)
p_k^U	Upper bound of the parent
p_k^L	Lower bound of the parent
δ_k	Deviation factor

d_i	Distance of i^{th} chromosome
x_i	Distance to the point on *x*-axis
y_i	Distance to the point on *y*-axis
R_{FN}	Rate of false negative
R_{FP}	Rate of false positive

Abbreviations

ANOVA	Analysis of Variance
BB	Blackberry
BLX-a	Blend Crossover
CP	Cherry Plum
CS	*Cymadocea Serrulata*
EU	Eucalyptus
FCB	Fuzzy Connective Based
FR	Fuzzy Recombination
GA	Genetic Algorithm
HO	*Halophia ovalis*
HS	*Halophia spinulosa*
HU	*Halodue univervis*
LM	Lemon
MOEA	Multiobjective Evolutionary Algorithm
NSGA	Nondominated Sorting Genetic Algorithm
NSGA-II	Nondominated Sorting Genetic Algorithm - II
nSSM	Normalized Spectral Similarity Measure
PSD	Probability of Spectral Discrimination
PWSD	Power of Spectral Discrimination
SAM	Spectral Angle Mapper
SBX	Simulated Binary Crossover
SCA	Spectral Correlation Angle
SCM	Spectral Correlation Measure
SI	*Syringodium isoetofolium*
SID	Spectral Infromation Divergence
SID-SAM	Spectral Infromation Divergence - Spectral Angle Mapper
VEGA	Vector Evaluated Genetic Algorithm
VNIR	Visible Near Infrared
ZM	*Zostera muelleri*

Chapter 1

Introduction

"A trick is nothing but a puzzle mankind came up with. If you use your head, you can uncover the logical answer."

Shinichi Kudo

1.1 Background

Periodic survey to identify plant in an ecosystem is important to monitor the health of ecosystem, which is measured by the change in the composition of species, for example, in the coastal environment (Rosenberg et al., 2004). Other than monitoring the health of the ecosystem, plant identification is also beneficial to early detection of the introduction of invasive species, such as *Undaria pinnatifida* or Wakame seaweed which was introduced by hull fouling to Australian coastal water (Hewitt et al., 1999).

The traditional way to identify plant species is by classifying its morphological characteristics, which can be complicated as the amount of material is insufficient for identification (Haider, 2011). Advances in computation allow the identification of plant species using conventional camera by shape recognition, i.e. using the difference in shape, colour and texture features are commonly used in identifying species (Hiremath and Pujari, 2007). Shape recognition, however, can be computationally expensive to apply and may not do very well in distorted images produced during survey (Davie et al., 2008). The use of only colour for identification purposes is appealing as it is less prone to geometric distortion. However, the conventional colour cameras do not capture sufficient colour information required to classify plant since they only capture three bands of visible light, Red-Green-Blue (RGB). A hyperspectral imager, either is a spectrometer or hyperspectral camera, on the other hand, captures more than 50 bands of electromagnetic radiation wavelength from ultraviolet to infrared. The narrow bandwidth and

continuous spectral information in a hyperspectral image allow for identification of different materials based on their spectral characteristics.

Hyperspectral imagers have been used in different areas of remote sensing application, such as geology (van der Meer, 2006), agriculture (Aasen et al., 2014), vegetation (Du et al., 2004) for decades. Reflectance spectra obtained from underwater hyperspectral imaging have been used as bio-taxonomic identification of corals (Kutser and Jupp, 2006, Botha et al., 2013), micro- and microalgae (Volent et al., 2009, Kutser and Metsamaa, 2006) and seagrass (Fyfe, 2001, Paringit et al., 2003). Fyfe (2001) sampled the spectral reflectance of three different seagrass species and found distinct spectral signatures of *Zostera capricorni, Halophia ovalis*, and *Posidonia australis* as shown in Figure 1. At 675 nm wavelength, *Z.capricorni* and *P. Australis* absorbed more light than *H. Ovalis*. Compared to the rest, *Z. capricorni* lacks intense green reflectance at 550 nm wavelength. The largest difference in the spectral reflectance of the three species takes place at the Near Infra-Red (NIR) region (700nm – 900 nm).

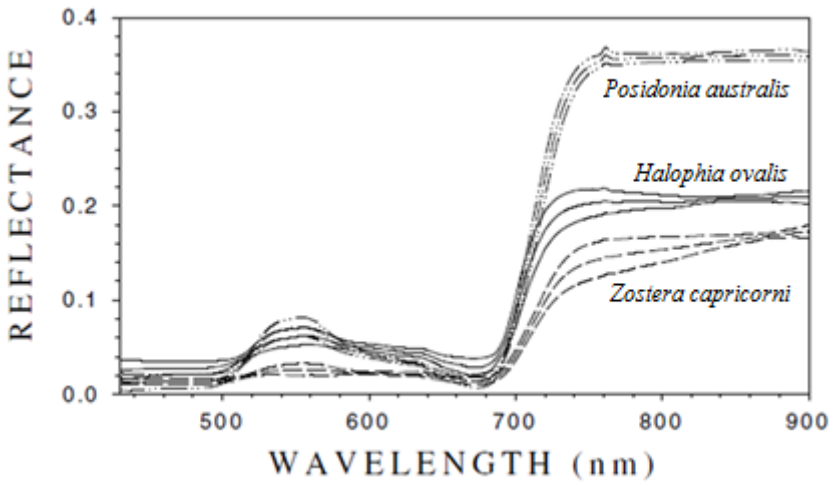


Figure 1 – Mean ± standard deviation of spectral reflectance of three south-eastern Australian seagrass species ((Fyfe, 2001)

1.2 Quantification of Spectral Information and Species Identification

A quantification, or measure, of the spectral information from different materials or biological species is required to facilitate classification. The so called spectral similarity measures effectively estimate the difference between spectra and have been used to distinguish minerals (van der Meer, 2006), corals (Botha et al., 2013) and vegetation and soils (Du et al., 2004). It is one of the most commonly used

methods in hyperspectral classification due to its simplicity as it does not require complex analysis and dimension reduction(Ma et al., 2016). The accuracy of the spectral similarity measure depends on its discrimination threshold, which varies for different material or biological species. The use of spectral similarity measure to classify a target species is carried out by the comparison of the spectra from the target species with the reference spectra in a spectral library. The measure assumes that the target is of a particular species if the similarity measure of target spectra and reference spectra of that particular species in the library is under certain discrimination threshold.

To date, no specific way has been identified to ascertain the definitive threshold value for discriminating terrestrial and underwater plants. The conventional way in remote sensing of minerals is to apply a user-defined or fixed value on a specific spectral similarity measure, such as 0.1 for Spectral Angle Mapper (SAM) (Shahriari et al., 2014). However, the conventional fixed threshold value does not necessarily provide a high capability to discriminate plants. Schwarz and Staenz (2001) introduced an adaptive threshold on the assumption of Gaussian distribution of the spectral angles. They concluded that the threshold value might be given by the mean minus three standard deviation. Shahriari et al. (2014) generalized this idea and proposed that the threshold value might be given by the mean minus *m* times the standard deviation, where *m* is a value ranging from one to three. The adaptive method shows better identification compared to the use of typical fixed value or user-defined method.

An optimum threshold determination is defined by the minimization of the probability of both false-positive and false-negative identification of species. False positive, also known as “false alarm”, refers to a similarity detection of otherwise dissimilar species. False negative, or “missed detection”, refers to similar species being recognized differently. A minimization problem of this type constitutes a multiobjective optimization which can be solved using the multiobjective variant of genetic algorithm.

1.3 Machine Learning Classifier

Another method that has been used to identify species is Support Vector Machine (SVM). It is a machine learning approach introduced by Vapnik (1995) to apply on supervised learning classification and regression problem. The advantage of this classifier is its ability to construct a hyperplane in the

infinite dimensional space, which gives the largest separation distance among the training data point. Since in principle, the larger separation distance the lower misclassification, SVM is a robust classification method. It has already been applied in the hyperspectral classification problem such as hyperspectral remote sensing classification (Gualtieri and Cromp, 1999), and underwater benthic classification (Bongiorno, 2014). Being considered as the state of art method of classification, SVM is used as a comparison benchmark in this thesis.

1.4 Research Aim and Objectives

The research in this thesis aims to optimize the classification of plant species using five different spectral similarity measures, namely Spectral Angle Mapper (SAM), Spectral Correlation Angle (SCA), Spectral Information Divergence (SID), SID-SAM and Normalized Spectral Similarity Measure (nSSM). As previously mentioned, the accurate classification requires the minimization of both false positive and false negative identification. The novel aspect of this research is about the formulation and the application of the Multiobjective Evolutionary Algorithm (MOEA), which is based on the genetic algorithm, to find the optimal threshold value of each of the five spectral similarity measures in identifying different seagrass and terrestrial species. Additionally, parameters contributing to the MOEA performance are also quantified. It is expected that the optimized threshold leads to more robust classification of seagrasses and terrestrial species using hyperspectral information.

1.5 Thesis Structure

The first part of Chapter 2 of this thesis presents a review of the formulation of different spectral similarity measures including the methods for quantifying their effectiveness in species identification. This is then followed by a description of the spectral data of the seagrass and terrestrial samples examined in the present research. Subsequently, the generic genetic algorithm is described and an introduction to the background of the MOEA and its variants is presented. This chapter also outlines the methodology for determining the best parameter value to optimize MOEA performance.

Chapter 3 discusses the results of the MOEA algorithm to determine the optimum threshold value for each spectral similarity measure. It is followed by a comparison of the misclassification rate from the adaptive threshold method, SVM, MOEA with fixed parameters, and MOEA with optimized parameters.

Chapter 4 provides the overall conclusions of the thesis and recommendations for future research.

Chapter 2

Methods

"Although this mirror can show a reflection, it cannot show you the truth."

Ai Haibara

2.1 Spectral Similarity Measures

The formulation of spectral similarity measures can be categorized as follows:

1. Deterministic, which includes Spectral Angle Mapper (SAM), Spectral Correlation Angle (SCA) and Normalized Spectral Similarity Measure (nSSM);
2. Stochastic, which includes Spectral Information Divergence (SID); and
3. Hybrid, such as SID-SAM.

The following sub-sections present a description of each measure:

2.1.1 Spectral Correlation Angle (SCA)

Van Der Meer and Bakker (1997) proposed a spectral correlation measure (SCM) of two spectral vectors, r_i and s_i, be given as a function of the Pearson's cross correlation (Pearson, 1901) given by:

$$SCM = \frac{n\sum_{l=1}^{n} r_{il}s_{il} - \sum_{l=1}^{n} r_{il}\sum_{l=1}^{n} s_{il}}{\sqrt{[n\sum_{l=1}^{n} {r_{il}}^2 - (\sum_{l=1}^{n} r_{il})^2][n\sum_{1}^{n} {s_{il}}^2 - (\sum_{l=1}^{n} s_{il})^2]}} \quad (1)$$

where *i* is the index of wavelength, and *n* is the number of spectral bands. Bajwa et al. (2004) then converted the correlation measure to angle measure which is shown as follows:

$$SCA(s_i, s_j) = cos^{-1}\left(\frac{SCM + 1}{2}\right) \quad (2)$$

The SCA takes into account the overall shape of the spectral signature and is relatively insensitive to the different brightness level, which is measured by the magnitude of the vector. Therefore, this measure

allows the matching of different albedo; however, a small shift in the position of the absorption band and shape on the spectral vector affects the SCA value. The two spectra are considered as similar if the SCA is zero and dissimilar if SCA is 1.57.

2.1.2 Spectral Angle Mapper (SAM)

Spectral angle mapper measures the angle of the dot product of two spectral vectors r_i and s_i:

$$SAM(r_i, s_i) = cos^{-1}\left(\frac{\sum_{l=1}^{n} r_{il} s_{il}}{\sqrt{\sum_{l=1}^{n} r_{il}^2}\sqrt{\sum_{l=1}^{n} s_{il}^2}}\right) \quad (3)$$

where *i* is the index of wavelength, and *n* is the number of bands. Similar to SCA, SAM is relatively insensitive to different brightness since the angle between two spectra vectors is independent from the length or magnitude of the vectors. The SAM goes from 0 when the spectra signatures are identical to 1.57 when they are completely different.

2.1.3 Normalized Spectral Similarity Measure (nSSM)

One of the most widely used metrics to measure the difference in albedo variation given by the spectral magnitude is the Euclidian distance (ED) (Keshava, 2004). Botha et al. (2013) introduced a Normalized Spectral Similarity Measures (nSSM) which combines the SAM that measures the difference in the spectral shape and a normalized Euclidian distance (NED) which accounts for the albedo variation. The nSSM is defined as:

$$nSSM = \sqrt{SAM^2 + NED^2} \quad (4)$$

where SAM is given in Eq. (3) and NED of the spectral signatures, r_i and s_i is given as:

$$NED = \sqrt{\frac{\sum_{l=1}^{n}(r_{il} - s_{il})^2}{\sum_{l=1}^{n}(r_{il} + s_{il})^2}} \quad (5)$$

where *i* is the index of wavelength, and *n* is the number of bands. The value of nSSM ranges from 0 when two spectra are identical to 1.86 when two spectra are different.

2.1.4 Spectral Information Divergence (SID)

Spectral information divergence is a stochastic measure which calculates the difference in the probability distribution of two spectra (Chang, 2003). The SID between two spectral signatures r_i and s_i is defined as:

$$SID(r_i, s_i) = D(r_i||s_i) + D(s_i||r_i) \tag{6}$$

where,

$$D(r_i||s_i) = \sum_{l=1}^{n} q_l\big(I_l(s_i) - I_l(r_i)\big) \tag{7}$$

and

$$D(s_i||r_i) = \sum_{l=1}^{n} p_l\big(I_l(r_i) - I_l(s_i)\big) \tag{8}$$

where *i* is the index of wavelength, and *n* is the number of spectral bands. $D(r_i||s_i)$ and $D(s_i||r_i)$ are called the relative entropy of s_i relative to r_i and the relative entropy of r_i relative to s_i, respectively. The probability vector $p = (p_1, p_2, \dots, p_l)$ and $q = (q_1, q_2, \dots, q_l)$ are given as $p_n = s_{il} / \sum_{l=1}^{n} s_{il}$ and $q_n = s_{jl} / \sum_{l=1}^{n} s_{jl}$. $I_l(r_{il})$ and $I_l(s_{il})$ are values of self-information measures given by $I_l(r_{il}) = -\log q_l$ and $I_l(s_{il}) = -\log p_l$, respectively.

The value of SID is equal to zero when the two spectral vectors are perfectly identical, and SID reaches infinity when the two spectral vectors are totally different.

2.1.5 SID-SAM

This hybrid method was first introduced by Du et al. (2004). It increases discriminability by making two similar spectra more similar and two different spectra more distinctive. There are two versions of this technique:

$$\text{SID-SAM(TAN)} = \text{SID}(r_i, s_i).\tan\big(\text{SAM}(r_i, s_i)\big) \tag{9}$$

and

$$\text{SID-SAM(SIN)} = \text{SID}(r_i, s_i).\sin\big(\text{SAM}(r_i, s_i)\big) \tag{10}$$

where r_i and s_i are the spectral signatures. The value of SID-SAM is equal to zero when the two spectral vectors are completely similar and infinity when the two spectral vectors are entirely distinct.

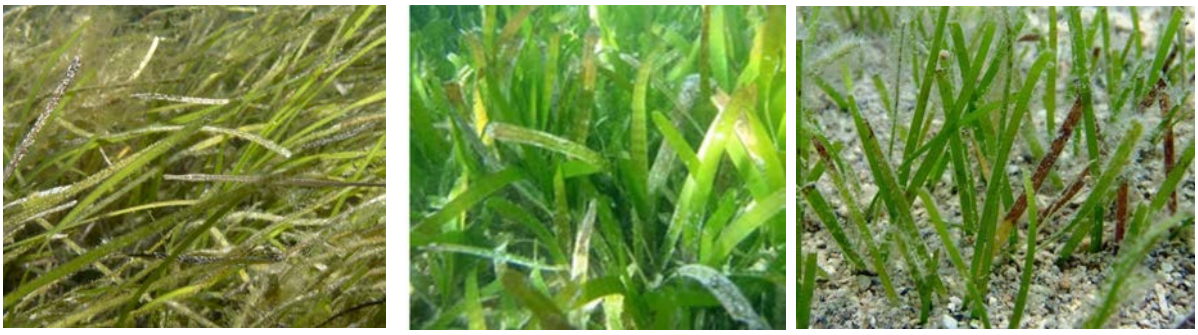
2.2 Spectral Data

2.2.1 Seagrass Samples

In principle, the purest spectra are those that are measured in-situ. However, measuring spectra reflectance underwater has several challenges due to limited light underwater and equipment capabilities, and since the scope of this thesis is only focusing on the classification. Therefore, the seagrass samples are only of those taken in air.

Seagrass samples examined in this thesis are *Halophia spinulosa* (Figure 2a), *Syringodium isoetifolium* (Figure 2b), *Halophia ovalis* (Figure 2c), *Zostera muelleri* (Figure 2d), *Halodue univervis* (Figure 2e), and *Cymadocea serrulata* (Figure 2f).

(a) (b) (c)



(d) (e) (f)

Figure 2- (a) Halophia spinulosa (Bryce, 2017), (b) Syringodium isoetifolium (Boisset, 2011), *(c) Halophia ovalis (*Boisset, *2015), (d)Zostera muelleri (Pocklington, 2011), (e) Cymadosea serrulata (McKenzie, 2007), (f) Halodue univervis(*Huisman, *2011)*

The spectral data of each sample is obtained from Roelfsema et al. (2016). They measured the reflectance spectra for wavelengths between 400 and 1050 nm in 1024 bands using ASD VNIR spectrometer. To account for the upwelling radiance (L_U) under the sunlight at the location of the measurement, the spectrometer optics was placed 15 cm above the target. To estimate the downwelling irradiance, the reflectance of a spectralon^{TM} panel, which is assumed to be 100% Lambertian surface, was measured at the same location within the same time period. The six types of seagrass species from different individuals were collected and measured on 16th January 2001 between 3 to 5 pm at Wangawallen banks, West North Stradbroke Island in Moreton Bay, Queensland (Figure 3).

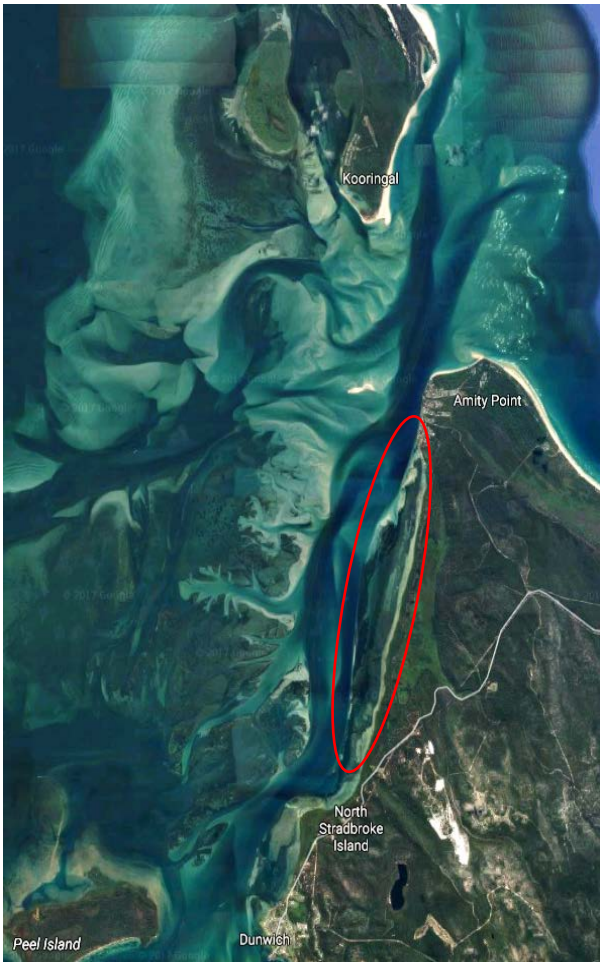


Figure 3 – Origin of seagrass sample (Wangawallen banks, west of North Stradbroke island) is circled in red

Ten samples of each seagrass were collected approximately 30 minutes before being measured and spread out over a black shade cloth. Spectral measurement was performed in air for all ten samples of

the six different types of seagrass. The radiance-reflectance signature was obtained by dividing the measured reflectance by the radiance reflected by the Spectralon (L_U/E_d).

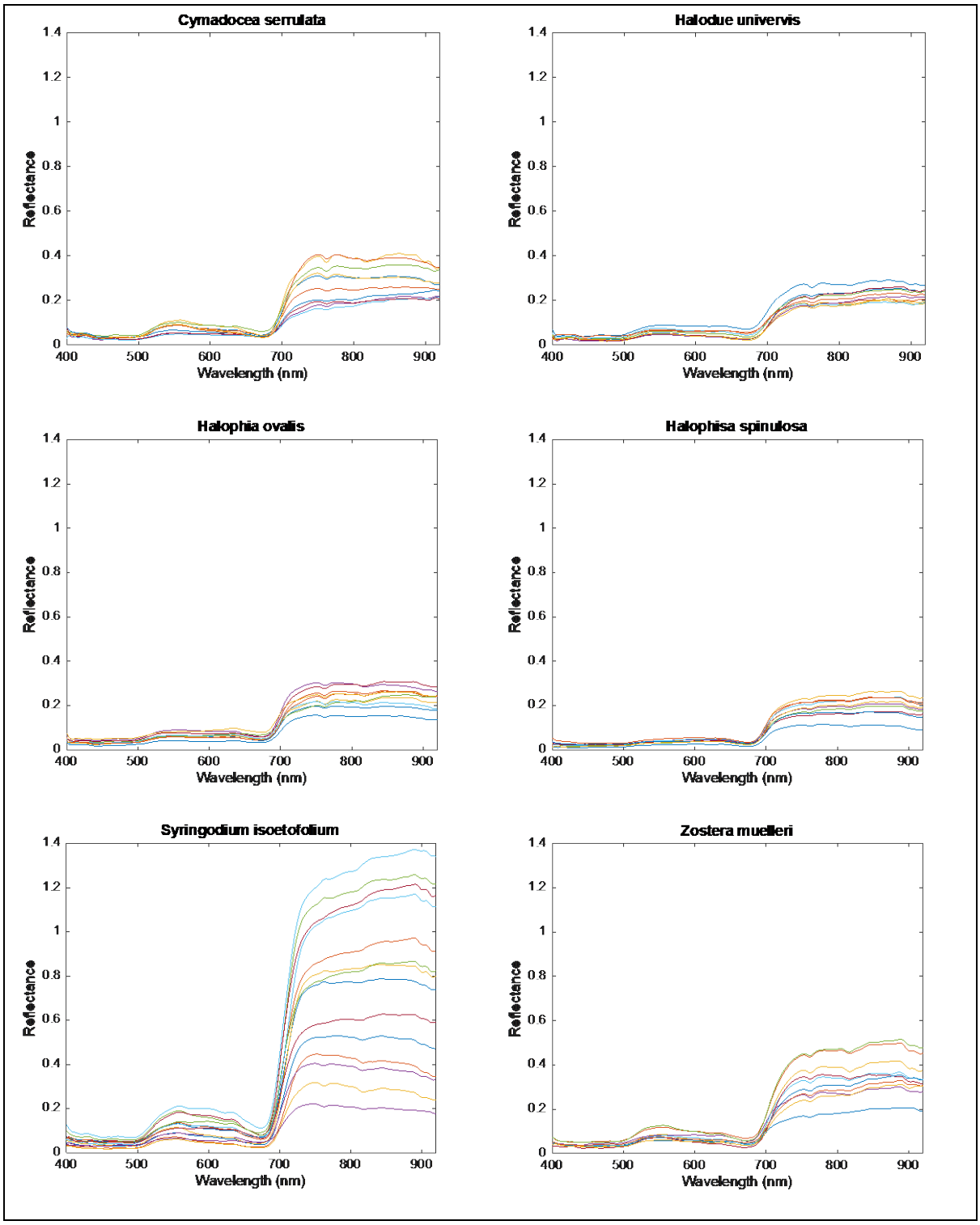


Figure 4 - Reflectance spectra of seagrass after smoothing and noise removal

Digital filtering and smoothing were subsequently applied to the measured reflectance data to remove any noise in the signal. Some noisy bands were also removed since it can distort the overall shape of

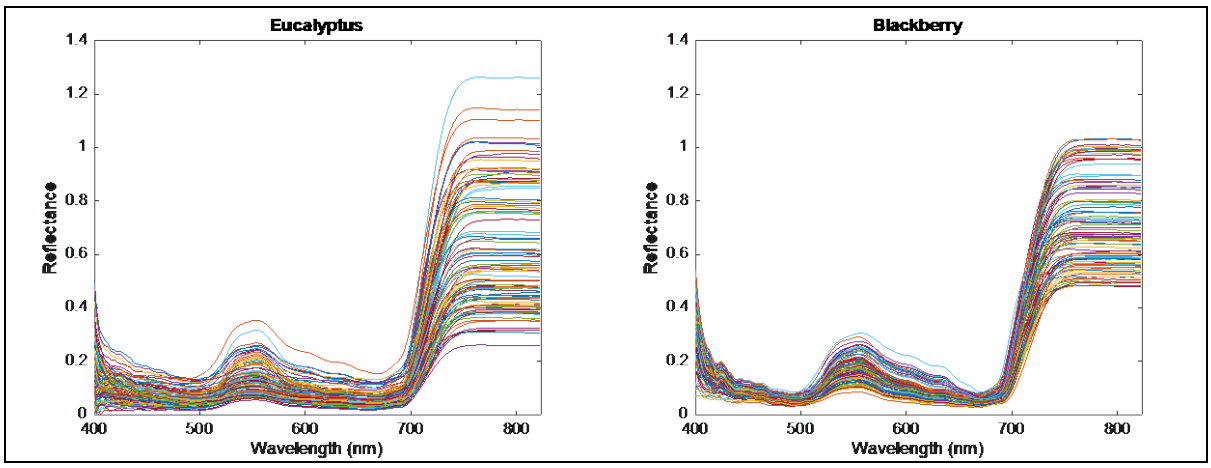
the signal and lead to lower correlation between spectral vectors. Figure 4 above shows the reflectance spectra after smoothing and removal of the noise.

The ten spectral data for each species are divided randomly into two sets: five spectra for training set and five spectra for evaluation set. The training set used for optimization purposes and the evaluation set is used to assess the performance of the optimized parameters of MOEA.

2.2.2 Terrestrial Plant Samples

Collection of spectral data of four species of terrestrial plants was carried out by the author of this thesis. The four species include Eucalyptus (*Eucalyptus globulus*), Lemon (*Citrus limon*), Blackberry (*Rubus fruticosus* agg.)**,** and Cherry plum (Prunus cerasifera). The measurement was performed using Ocean Optics STS-VIS spectrometer which measures the reflectance between 350 – 800 nm. The data acquisition was taken indoor using the OCEANVIEW program with Ocean Optics ECO-VIS light as light source and Spectralon as reference. Other source of light indoor was minimized during measurement to prevent interference on the reflectance.

A hundred-spectral data for each terrestrial plant were measured and divided randomly into two sets: fifty spectral data for training set and another fifty in evaluation set. Similar to the spectral data of seagrass samples, digital filtering and smoothing were also applied to those of the terrestrial plants. Bands below 400 nm were removed to prevent the distortion of the spectral reflectance shape. The processed spectra are shown in Figure 5.



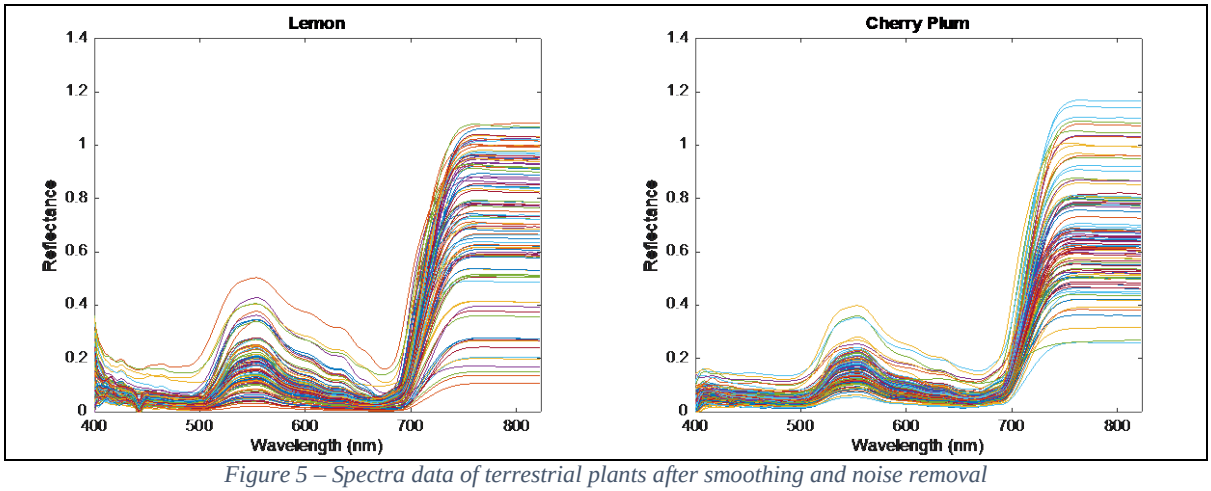


Figure 5 – Spectra data of terrestrial plants after smoothing and noise removal

2.3 Genetic Algorithm

Genetic algorithm (GA) is a directed random search algorithm that is inspired by natural evolution selection and genetics (Goldberg, 1988). It is, therefore, suited for applications in optimization or machine learning problems. Genetic algorithm was developed by John Holland and his collaborators at the University of Michigan (Holland, 1975).

The idea of GA is to evolve the population of solutions using operators, i.e. crossover and mutation which are based on the mechanics of natural selection and genetic variation. The crossover is the swapping of the parents' solutions (genes) to produce different offspring. Mutation is the mutating one solution (gene) in the parent chromosome to produce a different offspring. These operators aim to improve the fitness of the chromosomes over the generations. The fitness is defined as the closeness of the solutions to the optimum. The individuals that provide better solution (fitness) have greater probability to be preserved and reproduced in the next generation (Davis, 1991). The algorithm efficiently exploits the fitness evaluation of previous solution to predict a new solution with the expectation for an improved accuracy.

GA is considered more powerful than conventional optimization techniques as it allows for a direct manipulation of coding, a search from a population instead of a single point, a search via sampling and the use of a stochastic operator (Goldberg, 1988). A flow chart describing the generic form of the genetic algorithm for a single-objective optimization is shown in Figure 6.

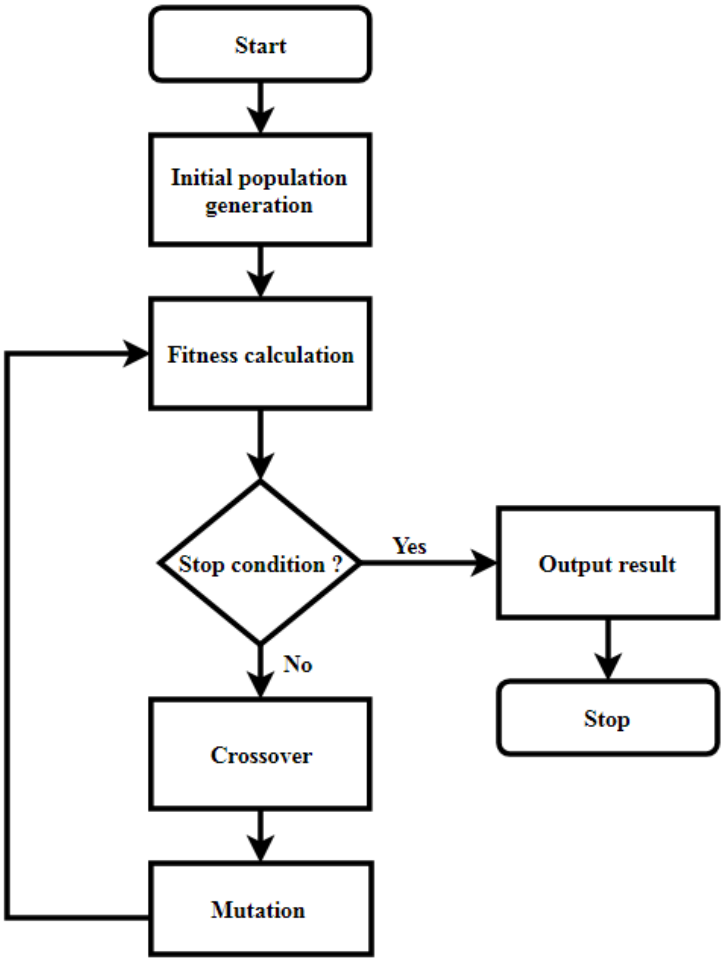


Figure 6 - Flow chart of single objective genetic algorithm

A modification to the generic form of GA needs to be applied for multiobjective optimization problems. There are two approaches available: preference-based and multiobjective evolutionary algorithms. Preference-based approaches were developed earlier based on the idea of converting multiple objective functions to a single objective optimization problem (Deb et al., 2002). It is usually done through a weighted sum of the multiple traits of the objective functions, or by converting the objectives into constraints (Horn et al., 1994). This method is simple to apply; however, assigning the appropriate weighting factors to obtain optimum solution is a challenge (Konak et al., 2006, Deb, 2001). It is known that the final solution is very sensitive to small changes in the weighting factor (Richardson et al., 1989). Some knowledge about the nature of the problem can provide guidance in selecting the appropriate weighting factors, however it may not be readily available (Zitzler, 1999).

The multi-objective evolutionary algorithm (MOEA), on the other hand, tries to determine a set of pareto optimal solutions. Pareto optimal solutions offer the best trade-off among the objectives. This approach is considered more robust than the preference based approaches since a large optimization search space can be handled and multiple alternative trade-offs can be generated within a single run (Zitzler, 1999, Deb et al., 2002).

2.4 Multiobjective Evolutionary Algorithm

The MOEA evolved from vector evaluated GA (VEGA) (Schaffer, 1985), non-dominated sorting GA (NSGA) (Srinivas and Deb, 1994) to NSGA-II (Deb et al., 2002). NSGA-II improves the performance of NSGA by replacing the sharing function with a crowd comparison approach. The new approach has better computational complexity. The flow chart of the NSGA-II algorithm used in this project is shown in Figure 7. At the start, the population chromosomes are generated from the spectral similarity measures on the training set from the collected spectral data. These chromosomes contain a set of genes which represents the threshold values of each measures. This is discussed further in Section 2.5. In the fitness evaluation, the threshold values in the chromosomes are compared against the comparison matrix i.e. a collection of value from spectral similarity measures of the similar and different spectra. The chromosome population is subsequently ranked using the fitness values which, in this case, are the probability of false positive and false negative of the classification (see Section 2.6). The ranked individuals are then sorted based on non-domination concept and the population is assigned with a crowding distance (see Section 2.7). Once the individuals are sorted and assigned with the crowding distance, a tournament selection is carried out using a crowded comparison operator (refer to Section 2.7.2). Afterwards, the chosen individual chromosomes undergo the crossover and mutation (Section 2.8). The resulting child/offspring population is combined with the current population and selection is implemented to choose the individuals for next generation. The next population is then sorted based on the non-domination. The subsequent generation is then filled by the better ranking fronts, then by non-crowded solutions until its size exceeds the current population size. These processes are iterated until the last generation. The output of this MOEA is the best chromosome, i.e. the chromosome that has the minimum probability of false negative and false positive.

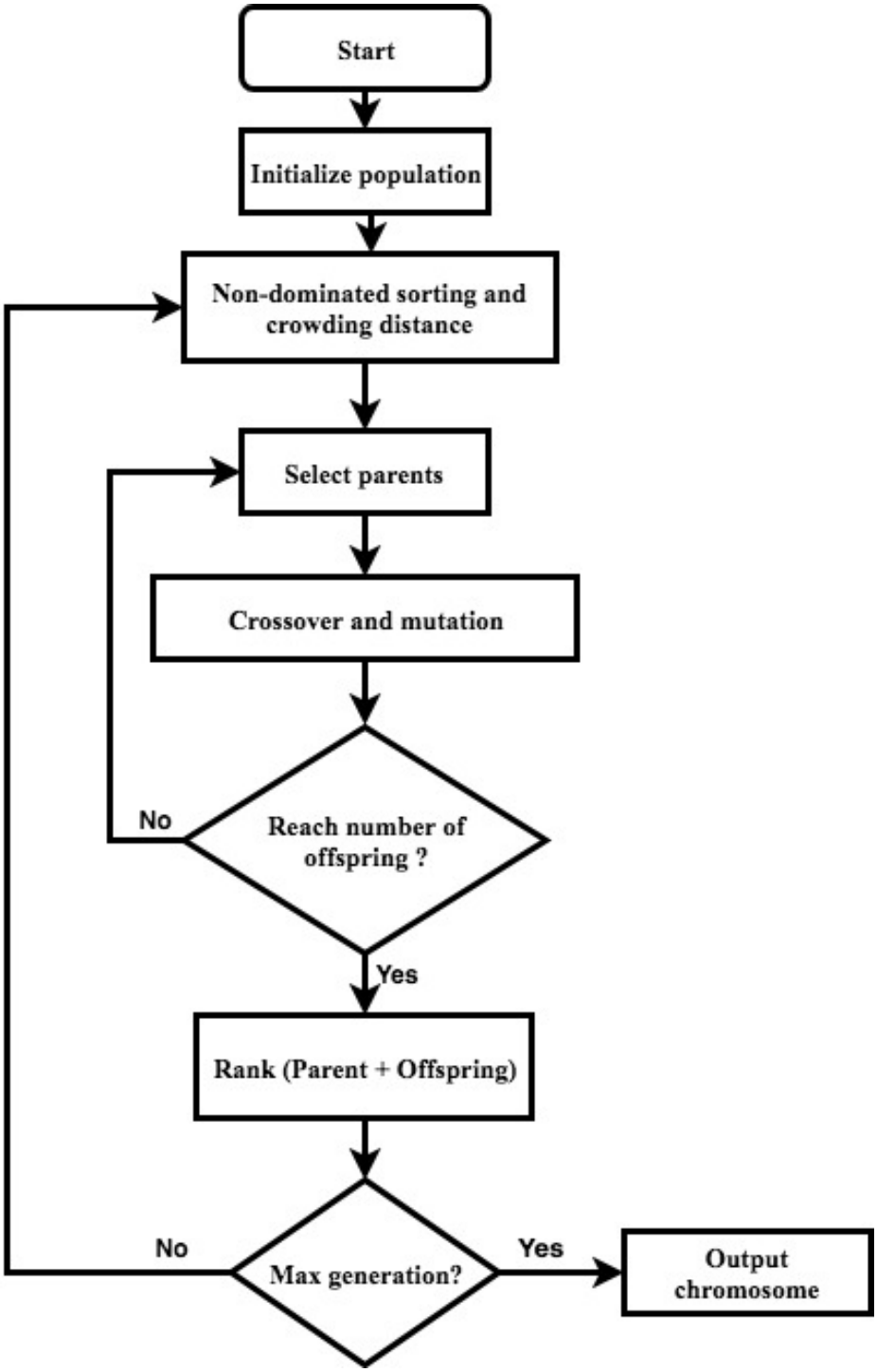


Figure 7- The flowchart multiobjective GA (NSGA-II)

2.5 Initialization

The first step in applying the MOEA is the initialization of chromosomes. A chromosome is a set of parameters that define the proposed solutions to the problem. Instead of randomly generated numbers to initialize chromosome, this research uses the spectral similarity measures from the spectra to obtain guided chromosome values.

The training spectra for each seagrass are assembled together in a 3-dimensional array of number of training spectra times number of species times number of bands. The spectral similarity measures, i.e. SAM, SCA, SID, SID-SAM and nSSM, are then applied to each spectral signature. The results are five matrices of spectral similarity measures each, which has the size of the square of number of training spectra times number of species. The first column corresponds to the permutation of comparison of two spectral signatures from one species, and the remaining columns correspond to the permutation of comparison of two different species. To make up a chromosome, the mean and standard deviation of each similarity measures are calculated from the first column of the matrices, and random numbers between the range of (mean – 1 times standard deviation) to (mean + 4 times standard deviation) are generated as genes. These steps for the initialization of the first set of chromosomes are illustrated in Figure 8 on the next page.

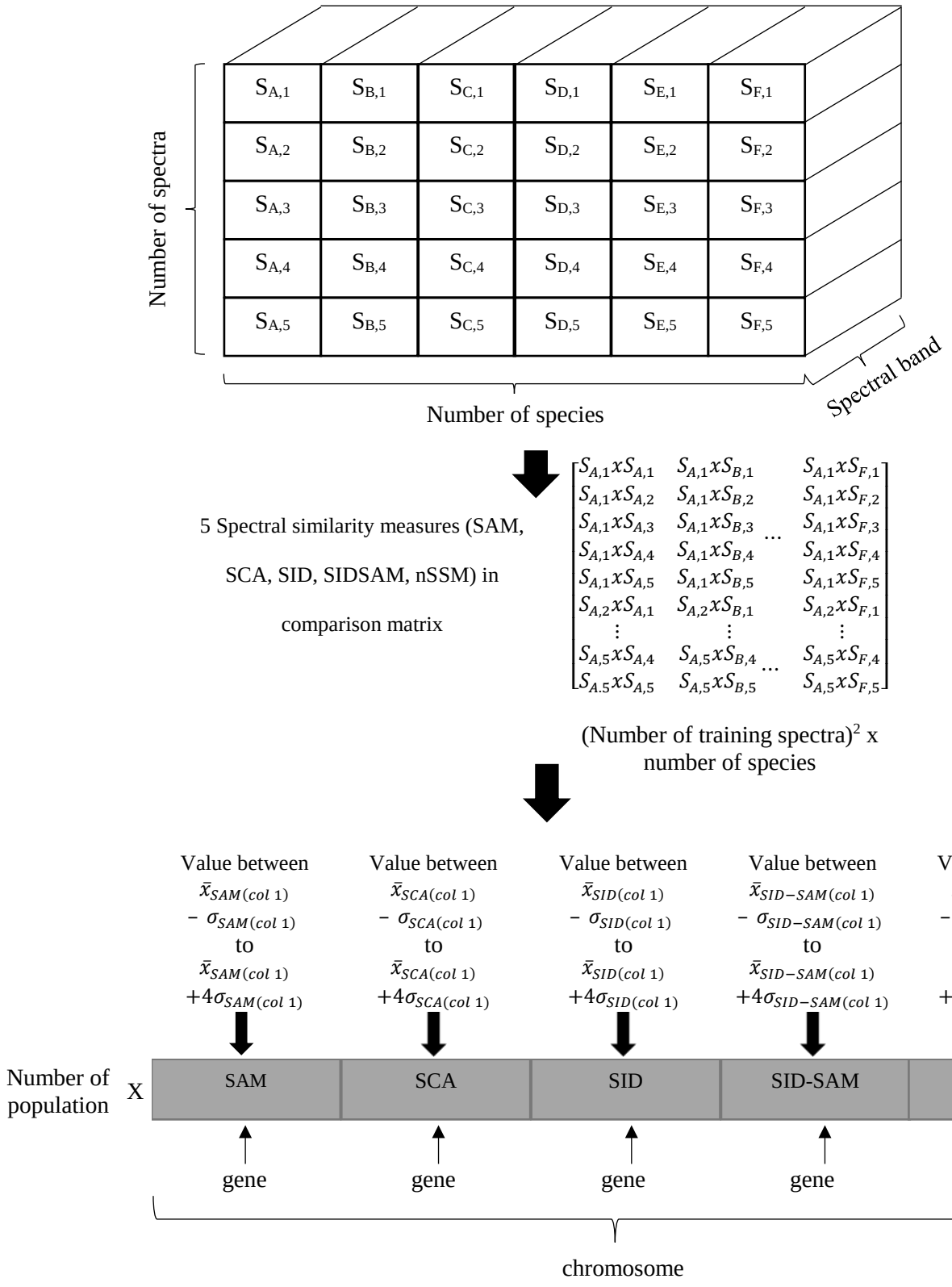


Figure 8 - Overview of chromosome initialization

2.6 Fitness Evaluation

The fitness of the chromosome is evaluated using by means of false positive and false negative classification of species in the training set as shown in the Table 1.

Table 1- Matrix for false positive and false negative classification

		Predicted		
		Positive	Negative	
Actual	Positive	True Positive (TP)	False Negative (FN)	Total Number of Positive NP = TP + FN
	Negative	False Positive (FP)	True Negative (TN)	Total Number of Negative NN = FP + TN

Figure 9 illustrates the process to determine the number of false positive and false negative classification. The value of each gene, which is a specific spectral similarity measure, is applied as a threshold value, to compare each element in the comparison matrix of one spectral similarity measure as shown in Figure 8. An element is considered as positive classification if it is less than the threshold value. The number of false positive and false negative classifications can then be ascertained from the location of the element since each row and column in the matrix corresponds to different or the same species.

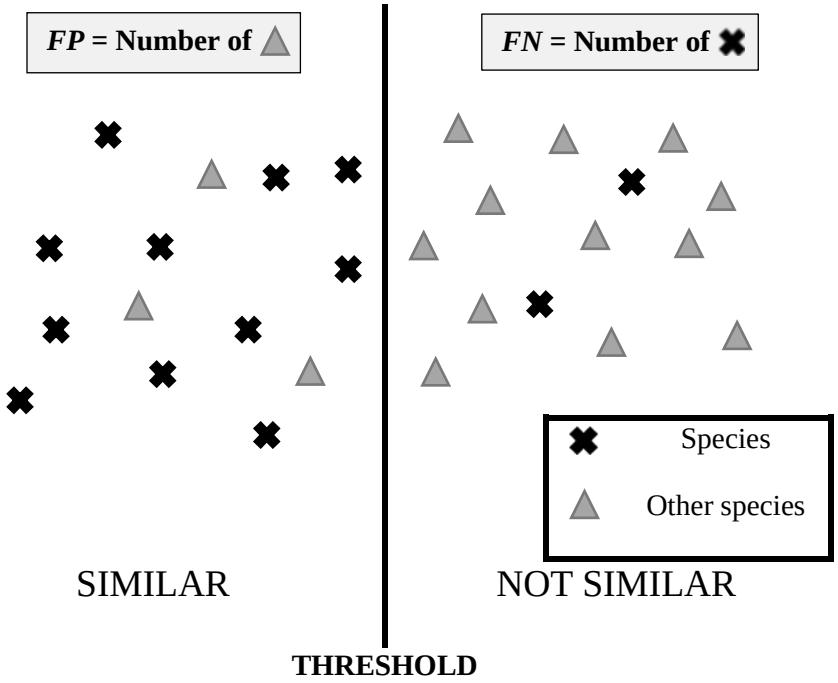


Figure 9 – Determining false positive and false negative classification

The objective function of this problem is minimization of the probability of false positive (P_{FP}) and minimization of probability of false negative (P_{FN}) which is shown below:

$$P_{FP} = FP/NN, \tag{11}$$

and

$$P_{FN} = FN/NP \tag{12}$$

where *FP* is the number of false positives, *FN* is the number of false negatives, *NP* is the number of positive classifications and *NN* is the number of negative classifications. P_{FP} and P_{FN} are computed for each gene since one gene represents the threshold value for a specific spectral similarity measures.

Since P_{FP} and P_{FN} for each gene, or threshold value, are statistically independent, the fitness function per chromosome can be written as:

$$P_{total\,FN} = \frac{\sum_{i=1}^{k} P_{FN,i}}{k} \tag{13}$$

$$P_{total\,FP} = \frac{\sum_{i=1}^{k} P_{FP,i}}{k} \tag{14}$$

where k is the number of genes, which is equal to five, i.e. the five different spectral similarity measures. Fitness value close to zero corresponds to the better optimization of the threshold.

2.7 Selection

The selection is the process to select which chromosomes are preserved for the next generation. The selection algorithm assigns two attributes to every chromosome i in the population: nondomination rank (i_{rank}) and crowding distance ($i_{distance}$).

The chromosomes in the next generation are chosen based on the rank after the non-dominated sorting. If two or more chromosomes have the same rank, then the individual from less crowded neighbourhood is chosen as illustrated in Figure 10. Further explanation on the non-dominated sort is given in Section 2.7.1 and on crowding distance in Section 2.7.2.

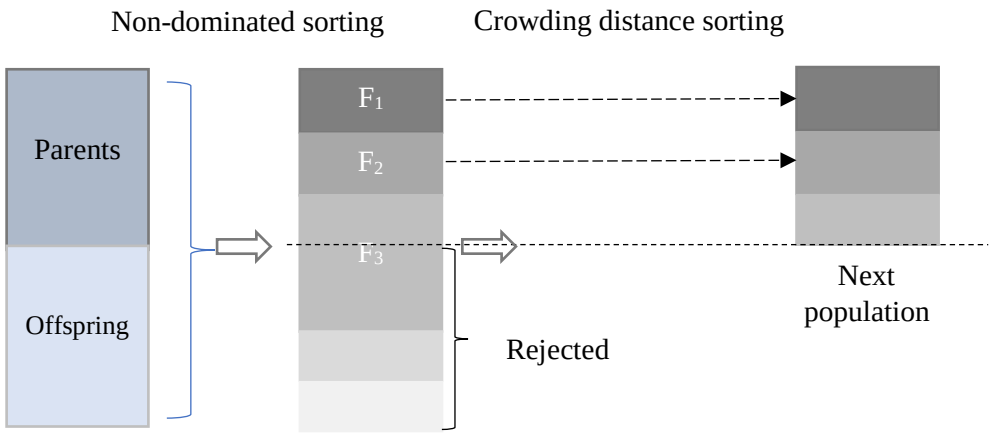


Figure 10 – Selection in NSGA-II (Deb et al., 2002)

2.7.1 Non-dominated Sort

Once the fitness values assigned to the chromosomes, they are sorted based on the non-dominated rule. Chromosome A is said to dominate B if A's fitness value of one objective is no worse than the B's and at least one of A's fitness value is better than the B (Seshadri, 2006) .The example is shown in Table 2.

Table 2 - Example of non-dominated sorting

	Fitness value	
	Objective 1	Objective 2
Chromosome A	0.1	0.2
Chromosome B	0.1	0.4

In Table 2, it is shown that chromosome A and chromosome B have same fitness value for objective 1. However, for objective 2, chromosome A has lower fitness value than chromosome B. In this case, chromosome A dominates chromosome B.

An illustration of the non-dominated sorting is shown in Figure 11 where F_1, F_2, F_3, and F_4 represent fronts 1, 2, 3, and 4, respectively. The front means a set of solutions ranked by the closeness to the optimum solution. In this case, front 1 has the highest rank since it is the closest to the optimum solution which is the origin of the coordinate system. The optimum solution in this case refers to the minimization of false negative classification and minimization of false positive classification. The front 1 is also called the non-dominated front as other fronts are further from the optimum solution. The idea of non-dominated sorting emphasizes good attributes and ensure maintaining healthy chromosome.

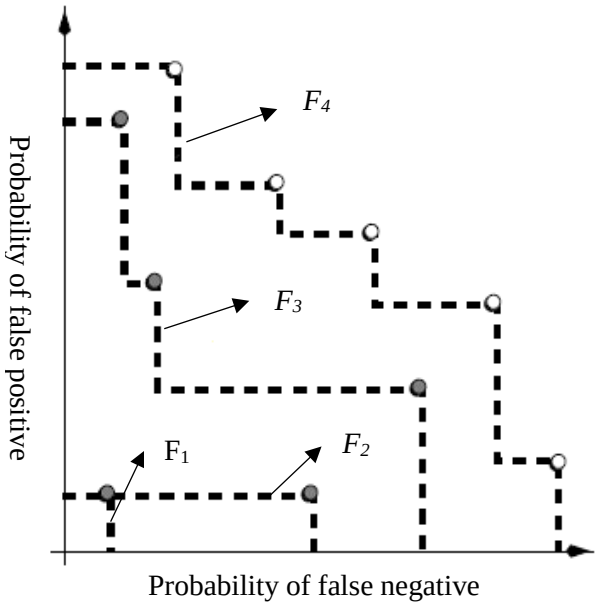


Figure 11 – Illustration of Non dominated sorting (Zitzler, 1999)

2.7.2 Crowding Distance

The purpose of crowding distance is to maintain the diversity of chromosomes in the population by selecting an individual from the less crowded neighbourhood. The level of crowding that surrounds a solution *i* is estimated by taking the average Euclidean distance between two solutions on either side of *i* on the same front. In Figure 12 the crowding distance of the *i*-th solution is the length of the longer side of rectangle with the solution *i-1* and *i+1* defining its diagonal. Thus, greater crowding distance implies less crowded neighbourhood.

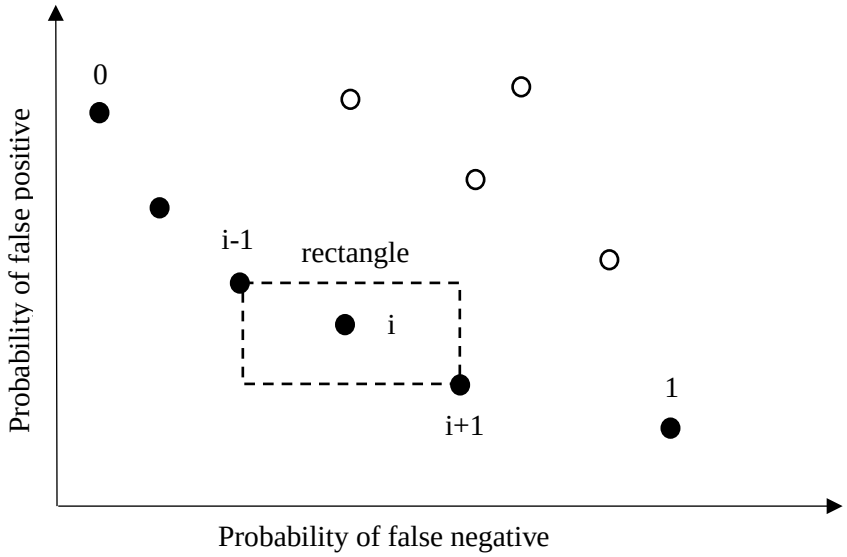


Figure 12 - Illustration of crowding distance calculation (Deb et al., 2000)

2.8 Crossover

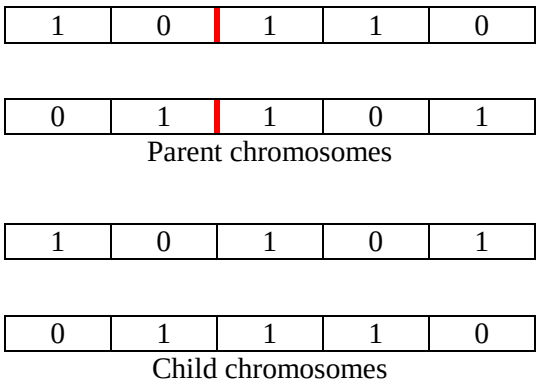


Figure 13 – Illustration of binary number crossover

The genetic algorithm uses crossover and mutation to create children that are different from the parents. The crossover happens when two parents exchange their genetic parts to make two children as shown in Figure 13 for binary coded crossover. The main purpose of crossover is to increase the diversity of the population since the selection process as described in Section 2.7 reduces diversity. Another purpose is to preserve the mean of the population since the crossover is not dependent on fitness values.

In a binary coded GA chromosomes are fixed in length and the crossover operation is easier to implement and visualize. However, in this research, the chromosome is coded in real number and it is more challenging to define decision vector in real number to create new offspring. There are many variants of real-number-coded crossover operators, such as simulated binary crossover (SBX) introduced by Agrawal (1995), blend crossover (BLX-a) by Goldberg (1991), fuzzy recombination (FR) by Voigt et al. (1995), fuzzy connective based (FCB) by Herrera et al. (1997), unfair crossover by Nomura and Miyoshi (1996). In the study by Beyer and Deb (2000) shows that SBX, BLX, and FR crossover operators have similar performance and act as a mean-preserving operator, whereas, other crossover operators do not preserve the population mean. Therefore, the SBX is chosen in this research.

2.8.1 Simulated Binary Crossover

Unlike other real value crossover, SBX uses a probability distribution which is similar, in principle, to binary coded chromosome crossover. SBX is based on a single point crossover in binary coded GA. The binary coded crossover is given as:

$$c_{1,k} = \frac{1}{2}\left[(1-\beta_k)p_{1,k} + (1+\beta_k)p_{2,k}\right] \tag{15}$$

$$c_{2,k} = \frac{1}{2}\left[(1-\beta_k)p_{1,k} + (1+\beta_k)p_{2,k}\right] \tag{16}$$

where $c_{i,k}$ is the i^{th} child with k^{th} component (k is gene element), $p_{i,k}$ is the selected parent chromosome and β_k is a random number with the density of:

$$p(\beta) = \frac{1}{2}(\eta_c + 1)\beta^{\eta_c}, \text{ if } 0 \leq \beta \leq 1 \tag{17}$$

$$p(\beta) = \frac{1}{2}(\eta_c + 1)\frac{1}{\beta^{\eta_c+2}}, \text{ if } \beta > 1. \quad (18)$$

This distribution, β, is obtained from uniformly sampled random number u between (0,1) and distribution index for crossover, η_c, given as

$$\beta(u) = (2u)^{\frac{1}{(\eta+1)}}, \text{ if } u \leq 0.5 \quad (19)$$

$$\beta(u) = \frac{1}{[2(1-u)]^{\frac{1}{(\eta+1)}}}, \text{ if } u > 0.5 \quad (20)$$

The distribution index, η_c, is kept constant through the process. It determines how close the offspring is to the parents. Large η_c means the offspring character is close to that of the parent and vice versa. This is illustrated in Figure 14.

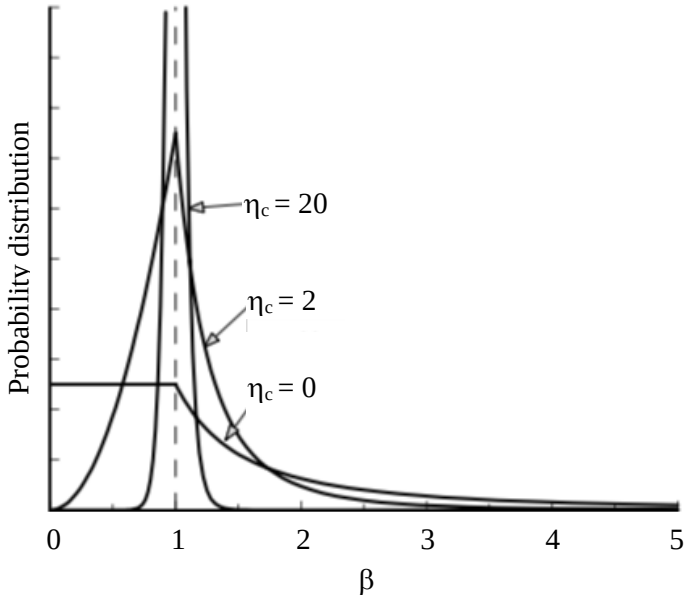


Figure 14 - Probability of expanding and contracting depending on the value of η_c (Agrawal, 1995)

An illustration of the probability distribution of the offspring character resulting from SBX is given in Figure 15. It shows that the probability of obtaining a child whose character is closer to that of the parents is higher than the probability of obtaining a child whose character is completely different from that of the parents.

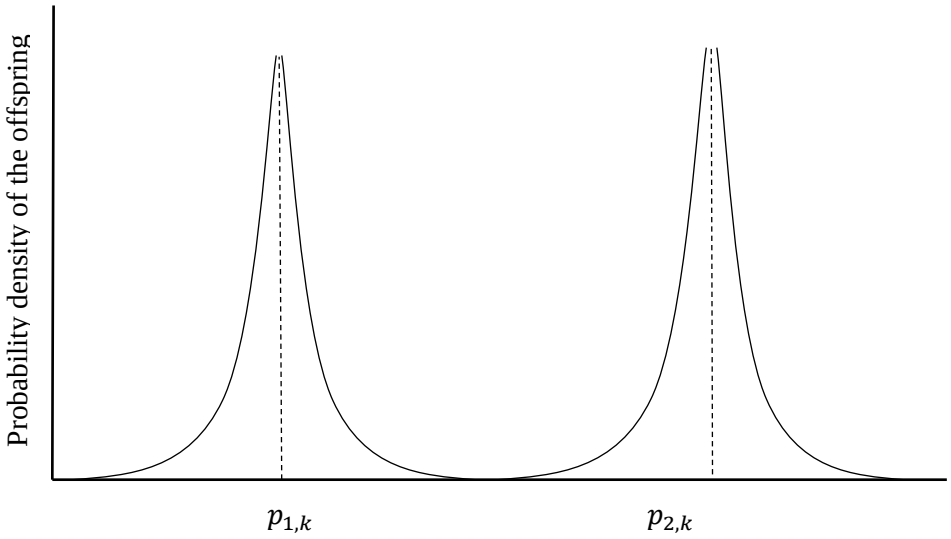


Figure 15 – The illustration of the simulated binary crossover with η_c = 20

2.9 Mutation

Mutation alters one or more genetic values in the chromosome from its initial state. Similar to the crossover operators, the aim of mutation is to increase the diversity of the solutions. In the case of binary-coded GA, a mutation refers to randomly flipping the value of a gene, i.e. from 1 to 0 or vice versa, as shown in Figure 16.

1	0	1	0	1

Parent chromosome

1	0	0	0	1

Child chromosome

Figure 16- Illustration of binary number mutation

Mutation of binary-coded chromosomes can have a negative effect on finding an optimum solution as it may disrupt the optimization process (Deb, 2001). In the case of chromosomes being coded in real number Gaussian and polynomial mutation may be used to ensure higher probability of having offspring whose character is close to that of the parent. Deb and Deb (2014) found that there is no difference in the resulting probability distribution between Gaussian and polynomial mutation, and hence polynomial mutation is adopted in the current work.

Polynomial mutation was first introduced by Deb and Goyal (1996) and is based on the polynomial distribution:

$$c_k = p_k + (p_k^U - p_k^L)\delta_k \tag{21}$$

where c_k is the offspring chromosome and p_k is the parent chromosome with p_k^U being upper bound, p_k^L is the lower bound of the parent component and δ_k is deviation factor calculated from polynomial distribution as follows:

$$\delta_k = (2r_k)^{\frac{1}{\eta_m+1}} - 1, \text{ if } r_k < 0.5 \tag{22}$$

$$\delta_k = 1 - [2(1 - r_k)]^{\frac{1}{\eta_m+1}}, \text{ if } r_k \geq 0.5 \tag{23}$$

where r_k is a randomly sampled number between 0 and 1 and η_m is mutation distribution index which is a user-defined parameter. This thesis uses η_m = 20 to maintain similar distribution spread of the offspring as crossover where η_m = 20 is used. An example of the probability distribution of the offspring character that results from polynomial mutation is illustrated in Figure 16.

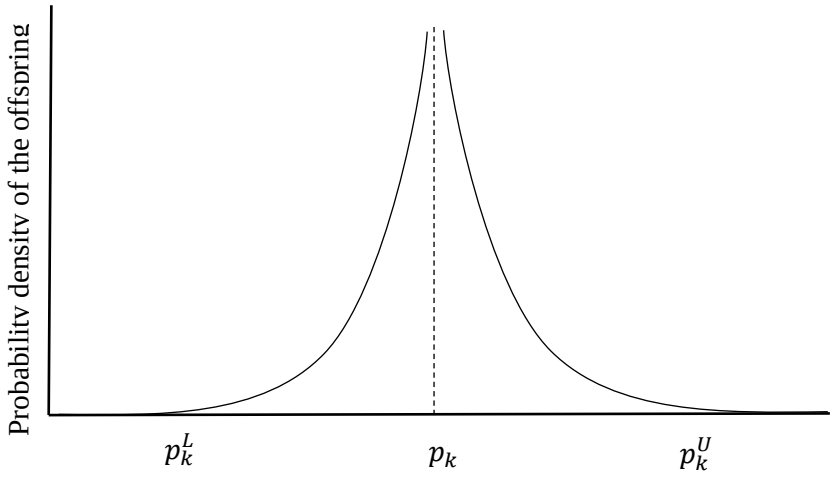
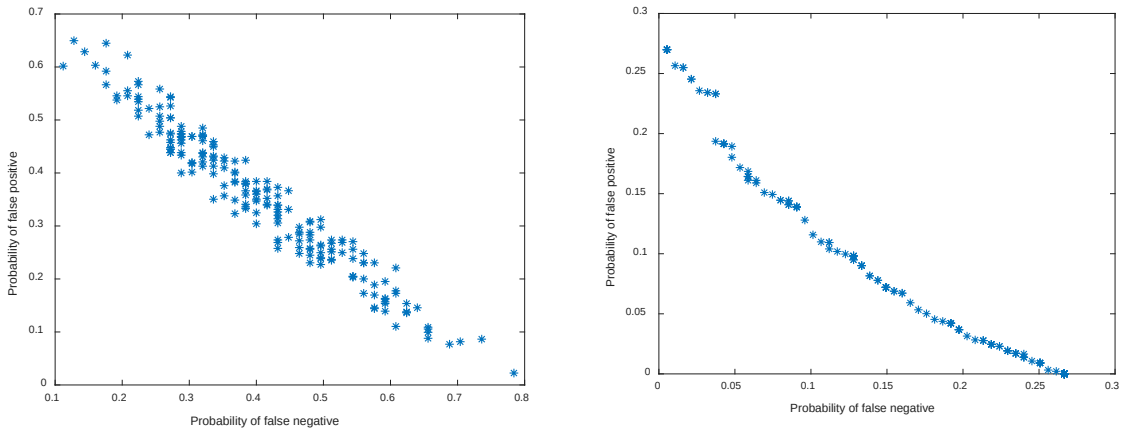


Figure 17 - The illustration of polynomial mutation

2.10 Output of Final Chromosome



(a) (b)

Figure 18 - The population in decision space (a) before NSGA-II, (b) after running 100th generation

As shown in Figure 18 the population of the chromosomes or solutions converges to a single front in the decision space after many generations. The population is the pareto optimal line where some solutions favour false positive and some solutions favour false positive. The final chromosome, i.e. the solution that minimizes both probability of false negative and probability of false positive, is determined by finding the minimum misclassification rate as the Euclidean distance to the origin of the coordinate system of the decision space as shown in Figure 18.

$$misclassification\ rate = \sqrt{{R_{FP}}^2 + {R_{FN}}^2} \qquad (24)$$

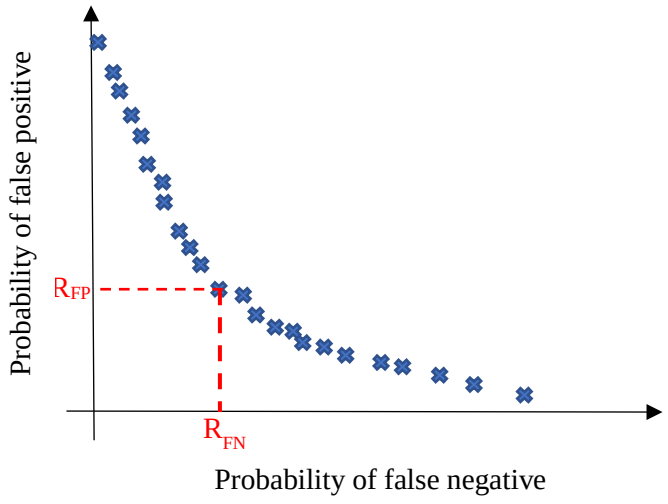


Figure 19 - Illustration of the distance measurement

2.11 Application to Evaluation Set and Parametric Studies

After the optimum parameters of the MOEA are identified using the training set, the parameters are then applied to the classification of species in the evaluation set.

In this thesis, different parameters were examined to compare the performance of unoptimized MOEA and that of optimized MOEA. The unoptimized MOEA is based on the default values described in Section 2.11.1, whereas, the optimized MOEA uses the best estimated values from a series of parametric study.

2.11.1 Unoptimized MOEA

The unoptimized MOEA uses the default values in NSGA-II as recommended in (Deb et al., 2000) Seshadri (2006) such as crossover rate of 0.9 and mutation rate of 0.1. The parameters for unoptimized MOEA are presented in Table 3 below.

Table 3 – Parameters for unoptimized MOEA

Variable representation	Real number
Decision variable	5 (SAM, SCA, SID, SID-SAM, nSSM)
Population size	100 chromosomes
Number of generation	100 generations
Selection	Non-dominated sorting and crowding distance
Crossover operator	Binary simulated crossover with rate of 0.9
Mutation operator	Polynomial mutation with rate of 0.1
Performance measure	The optimum of the final population i.e. the shortest distance to the origin (see Eq. (24))

2.11.2 Optimized MOEA

The behaviour and performance of the genetic algorithm is depends on the correct fine tuning of its parameters (Rojas et al., 2002). The purpose of parametric studies in this thesis is to evaluate the result

of MOEA when the parameters are varied and to find out which parameters significantly affect its performance. While there are many factors that may contribute to the performance of the MOEA, such as different type of crossover and mutation formulation, the following four parameters are considered to affect the performance of MOEA the most:

1. Number of generation;
2. Initialization of chromosomes;
3. Crossover rate; and
4. Mutation rate.

The first study covers the number of generation, which is considered one of significant factors affecting the performance of GA. If the number of generation is not sufficient, the search is terminated when optimum solution has not been reached. The study runs for 30 times on the training set which is chosen randomly from the spectral data and the result is taken as the average of the run. The optimum number is said to have reached convergence when there is 0.01% or less change in the solution. The study runs on initialization range of ($\bar{x} - \sigma$ to $\bar{x} + 4\sigma$), crossover rate of 0.9, and mutation rate of 0.1. Then, the optimum number of generation is then used for the next study.

In his thesis on statistical analysis of GA, Czarn (2008) argued that the initialization chromosome seeds can be a statistically significant factor in the MOEA result. Since initialization chromosome may interact with crossover and/or mutation rate, this parametric study uses 3-way ANOVA to test their significance. The study is performed with 4 different level initialization seeds, 4 different level of crossover rate, and 4 different level of mutation rate. So, in total, there are $4^3 = 64$ combinations. In running the ANOVA test, two types of error may be committed, i.e. type I and type II error. Type I error is when null hypothesis is rejected when it is true. If null hypothesis is not rejected when it is false, then type II error has been committed. The value of the probability of type I error α, is set at 0.05 with 95% confidence level in this ANOVA test. The probability of type II error β, is set at 0.2 which means a power of 80% rejecting the null hypothesis when it is false, which is moderate departure from the null hypothesis. The value of β is related to the sample size. Too small sample size will generally produce insignificant result, whereas too large sample size may be hard to analyze and waste resources. The

power calculations were carried out using GPOWER (Erdfelder et al., 1996) assuming that the effect size is moderate yielded at least 7 runs per sample group.

To minimize random variability and type II error in each combination, there are 8 runs for every combination. The initial chromosome is generated randomly within a certain range of ($\bar{x} - \sigma$ to $\bar{x} + m\sigma$), where *m* is an integer of 3,4,5, and 6. The crossover rate runs from 0.25 to 1 with interval of 0.25, and the mutation rate runs from 0.1 to 0.4. The ANOVA uses the 95% confidence level (α = 0.05), which means the null-hypothesis is rejected if p-value less than 0.05. The assumptions of ANOVA, which are normality of the residuals, and the non-correlation of the residuals versus the fitted/predicted values, are also checked to ascertain whether the ANOVA conclusion is valid.

Chapter 3

Results

"Life is short so it will be appreciated. We live to our fullest due to that"

Heiji Hattori

3.1 Result of Parametric Studies on MOEA Optimization

This section presents the results of the parametric studies described in Section 2.11 on optimization of MOEA, which focuses on four different parameters, i.e. number of generations (Section 3.1.1) , crossover rate, mutation rate, and initialization of chromosome (Section 3.1.2).

3.1.1 Number of Generations

As explained in Section 2.11.2, the number of generations is being investigated since MOEA requires sufficient generations to reach convergence before it is being terminated. The investigation results for determining the optimum number of generation are presented in Figure 20 for the six types of seagrass, and in Figure 21 for the four types of terrestrial plants.

The result in Figure 20 shows that MOEA converges at different number of generations on each seagrass species. *Cymadocea serrulata* converges at 300 generation, *Halodue univervis* converges at 230 generation, *Halophia ovalis* converges at 310 generations, *Halophia spinulosa* converges at 590 generations, *Syringodium isoetofolium* converges at 360 generations, and *Zostera muelleri* converges at 270 generations.

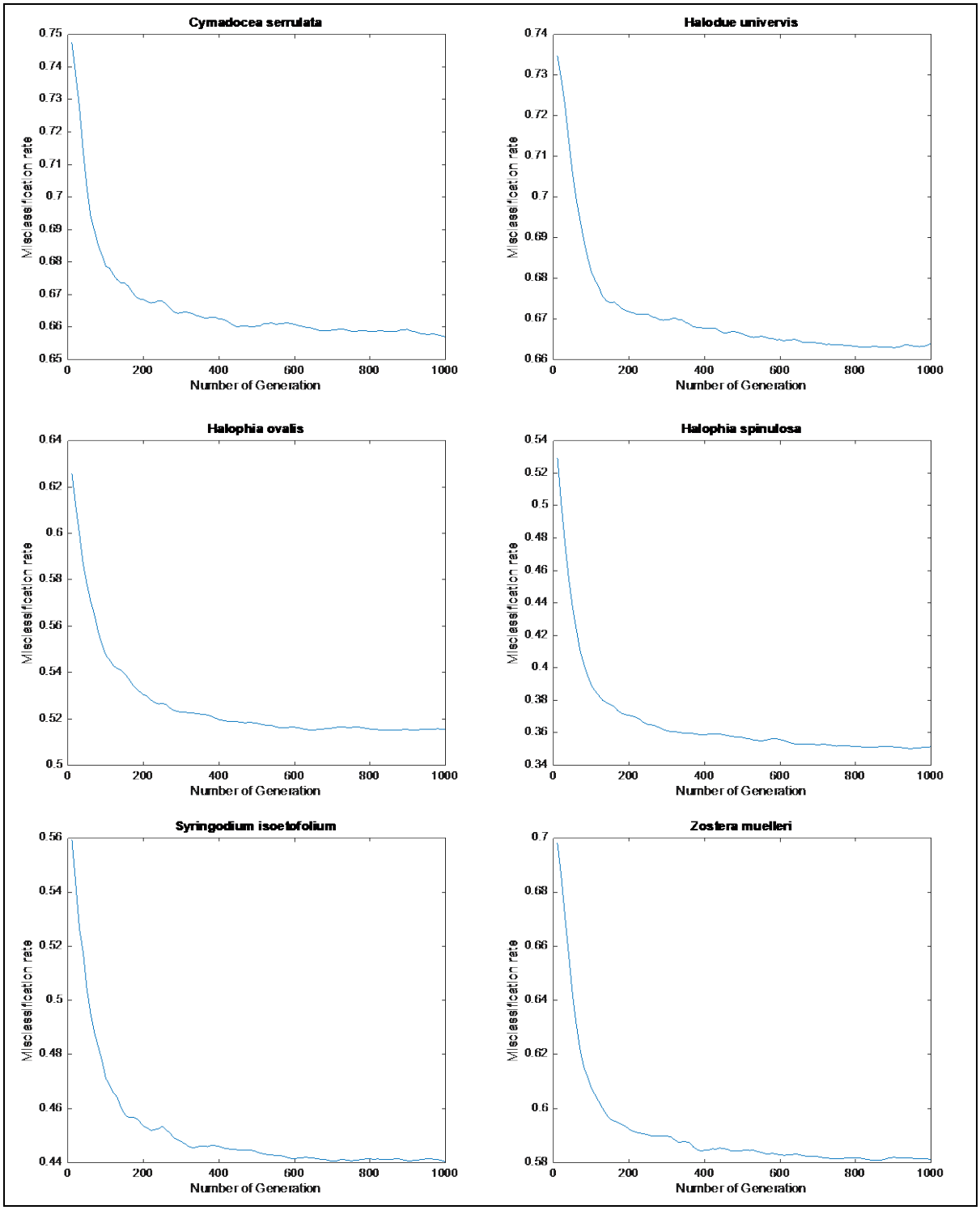


Figure 20 – Effects of different number of generation to misclassification rate (seagrass)

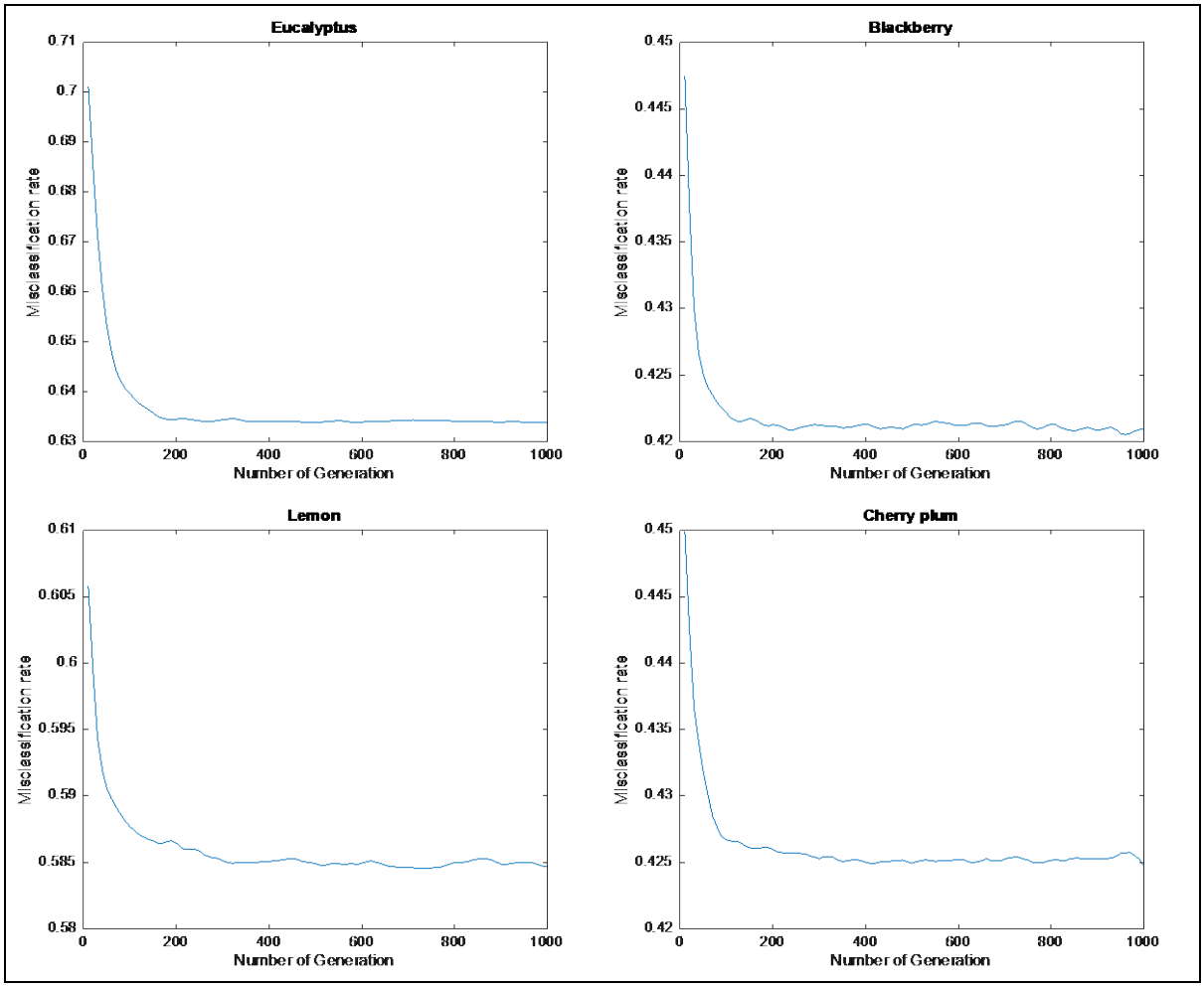


Figure 21 - Effects of different number of generation to misclassification rate (terrestrial plants)

The result presented in Figure 21 shows that the terrestrial plants converge relatively faster than the seagrass species. The difference in the number of generations required for convergence might be attributed to the low variation in the terrestrial plant's spectra reflectance. The eucalyptus converges after 190 generations, blackberry converges after 300 generations, lemon converges after 170 generations and cherry plum after 120 generations. These computed optimum numbers of generations are used in the subsequent study of finding optimum initialization of chromosome, crossover and mutation rate in Section 3.1.2.

3.1.2 Initialization of Chromosome, Crossover and Mutation Rate

The results of the parametric studies on six types of seagrass and four types of terrestrial plants are shown in Sections 3.1.2.1 through 3.1.2.10. Since the ANOVA is based on 95% confidence level, if p-value (Prob > F) of certain parameter is less than 0.05, then the parameter is said to be statistically significant.

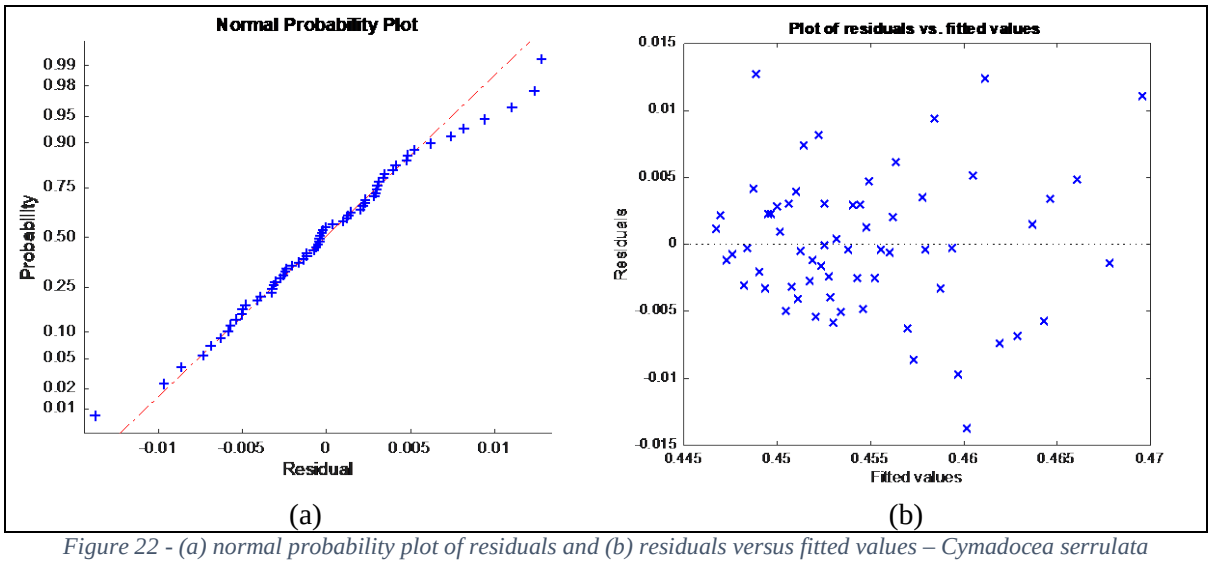
3.1.2.1 Cymadocea serrulata

The result of 3-way ANOVA for *Cymadocea serrulata* is shown in Table 4.

Table 4 – Result of 3-way ANOVA – Cymadocea serrulata

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	0.00087	3	0.00029	11.6	**0**
Mutation	0.00038	3	0.00013	4.99	**0.007**
Range	0.00044	3	0.00015	5.81	**0.0034**
Crossover * Mutation	0.00023	9	0.00003	1	0.461
Crossover* Range	0.0006	9	0.00007	2.65	**0.0242**
Mutation*Range	0.00031	9	0.00003	1.36	0.2559
Error	0.00068	27	0.00003		
Total	0.00349	63			

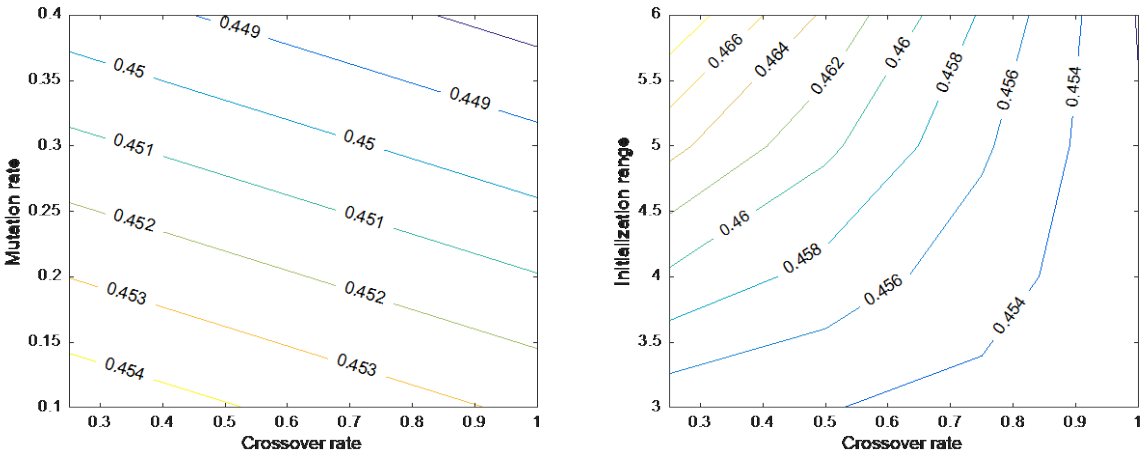
The result presented in Table 4 implies that the crossover rate, mutation rate, and initialization of chromosome are statistically significant. In addition, the ANOVA also finds that the interaction between crossover and range is significant. To ensure that this conclusion is correct, the study on normality of residuals and autocorrelation of residuals is presented in Figure 22.



(a) (b)

Figure 22 - (a) normal probability plot of residuals and (b) residuals versus fitted values – Cymadocea serrulata

At first sight, the normal probability plot for residuals as shown in Figure 22(a) appears slightly skewed. However, a normality check using Jarque-Bera test fulfils the assumption of normality of the residuals for ANOVA. The residuals plot also appears to indicate normality without discernible trend, therefore there are no reason to suspect the validity of ANOVA. Using the result of ANOVA, the response surface contour plots are generated by the regression model and shown in Figure 23.



(a) Contour plot with initialization range of 3 (b) Contour plot with mutation rate of 0.1

Figure 23 – Contour plot of misclassification rate – Cymadocea serrulata

The surface response contour plot in Figure 23(a) and (b) suggests that low misclassification rate favours high crossover rate, high mutation rate, and low initialization range for initial chromosome.

3.1.2.2 Halodue univervis

The result of three-way ANOVA for *Halodue univervis* is presented below in Table 5.

Table 5 – Result of 3-way ANOVA – Halodue univervis

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	0.00061	3	0.0002	40.12	**0**
Mutation	0.00005	3	0.00002	3.45	**0.0304**
Range	0.00005	3	0.00002	3.17	**0.0402**
Crossover * Mutation	0.0001	9	0.00001	2.09	0.067
Crossover* Range	0.00003	9	0.00001	0.62	0.7726
Mutation*Range	0.0008	9	0.00001	1.74	0.1291
Error	0.00014	27	0.00001		
Total	0.00105	63			

Table 5 shows that the crossover rate is the most significant parameter, whereas, mutation rate and range are less statistically significant. The investigation for the validity of ANOVA assumptions is displayed in Figure 24.

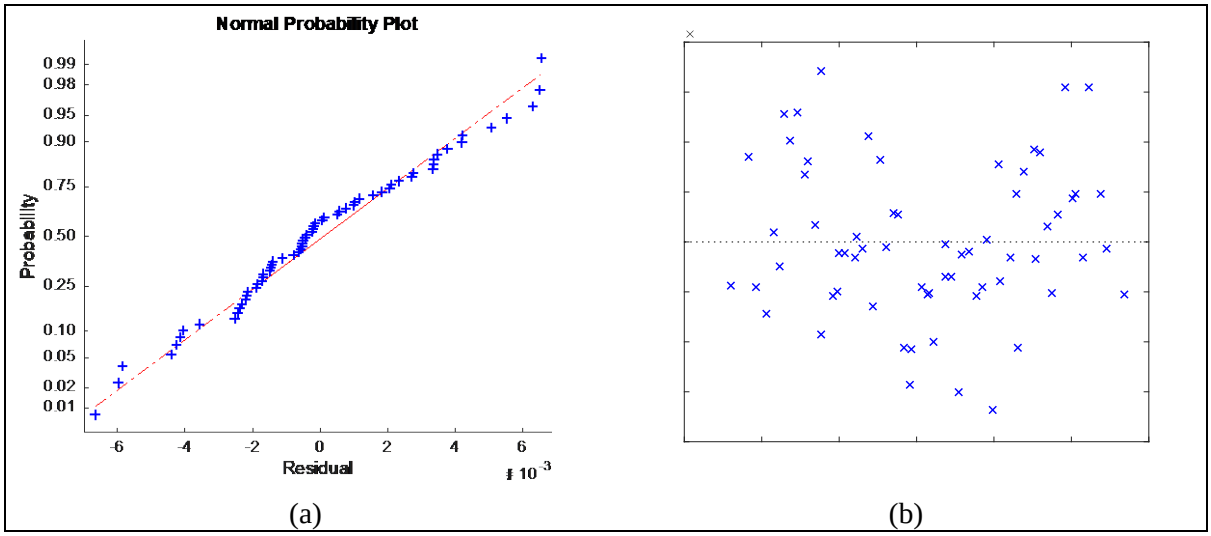


Figure 24 – (a) normal probability plot of residuals and (b) residuals versus fitted values – Halodue univervis

The normal probability plot of residuals in Figure 24(a) and plot of residuals versus fitted values appear to be satisfactory. Hence, there can be ascertained that the conclusion of ANOVA is valid. Then, the response surface contour plots (Figure 25) are generated using regression model to analyse the parameters.

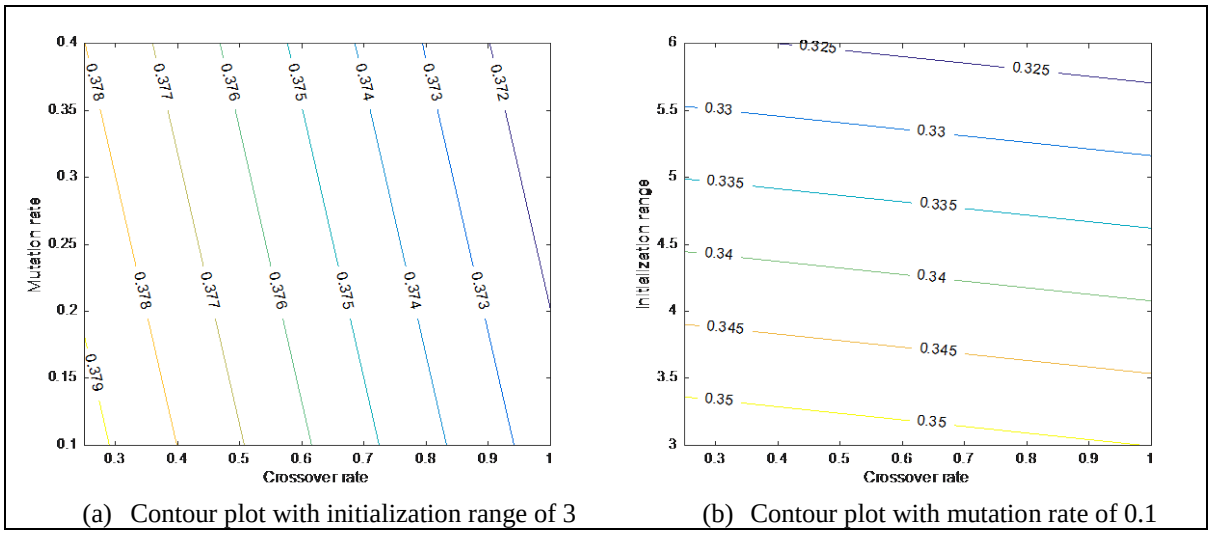


Figure 25 - Contour plot of misclassification rate – Halodue univervis

The response surface contour plots in Figure 25 (a) and (b) show that the minimization of misclassification rate requires high crossover rate, high mutation rate, and high initialization range.

3.1.2.3 Halophia ovalis

The outcome of three-way ANOVA test for *Halophia ovalis* is presented in Table 6.

Table 6 - Result of 3-way ANOVA – Halophia ovalis

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	0.00276	3	0.00092	92.87	**0**
Mutation	0.00012	3	0.00004	4.08	**0.0164**
Range	0.00073	3	0.00024	24.65	**0**
Crossover * Mutation	0.00011	9	0.00001	1.24	0.3155
Crossover* Range	0.00061	9	0.00007	6.83	**0**
Mutation*Range	0.00014	9	0.00002	1.53	0.1879
Error	0.00027	27	0.00001		
Total	0.00473	63			

The result of ANOVA indicates that the crossover rate, mutation rate, and initialization range are statistically significant to the misclassification rate. There is also significant interaction between the crossover rate and initialization range. To confirm if the ANOVA is valid, the normality analysis on the residuals and the non-correlation among the variables on the residuals are presented below in Figure 26.

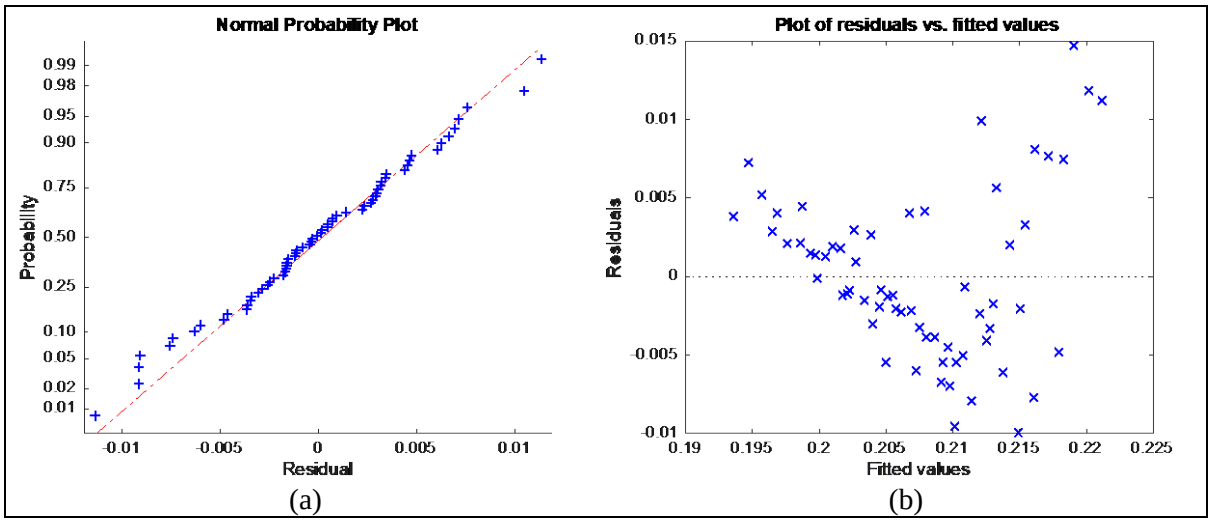
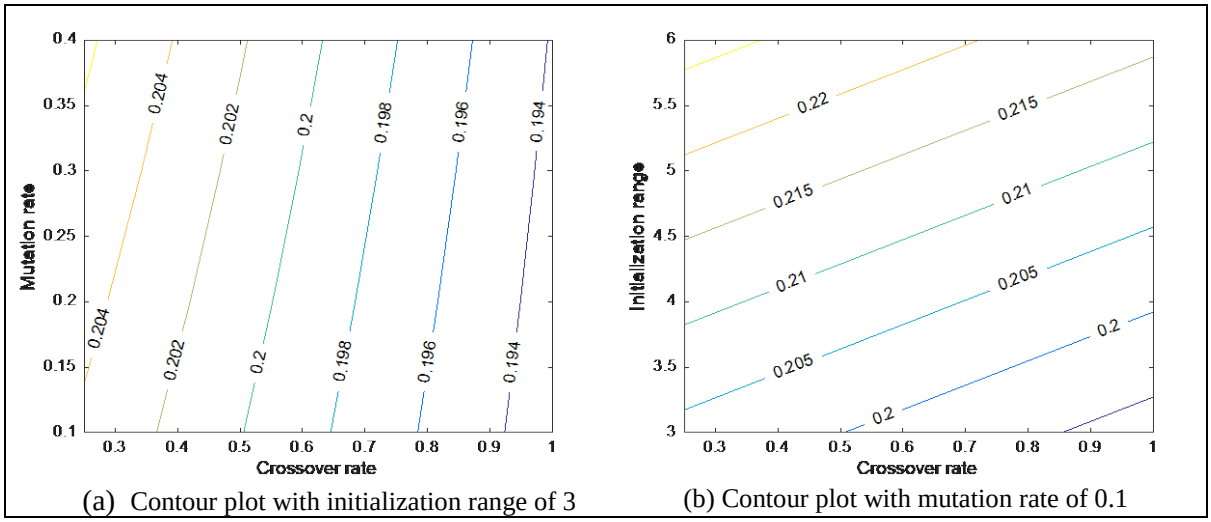


Figure 26 – (a) normal probability plot of residuals and (b) residuals versus fitted values – Halophia ovalis

The normal probability plot of residuals in Figure 26(a) appears to be satisfactory. The residuals versus fitted values in Figure 26(b) shows no visible pattern that there is no correlation between the variables in the residuals. Therefore, the assumptions of ANOVA are valid. Then, the surface response plot is generated using regression model in Figure 27.



(a) Contour plot with initialization range of 3 (b) Contour plot with mutation rate of 0.1

Figure 27 - Contour plot of misclassification rate – Halophia ovalis

The response surface contour plot shown in Figure 27 displays that low misclassification favours high crossover rate, low mutation rate, and low initialization range.

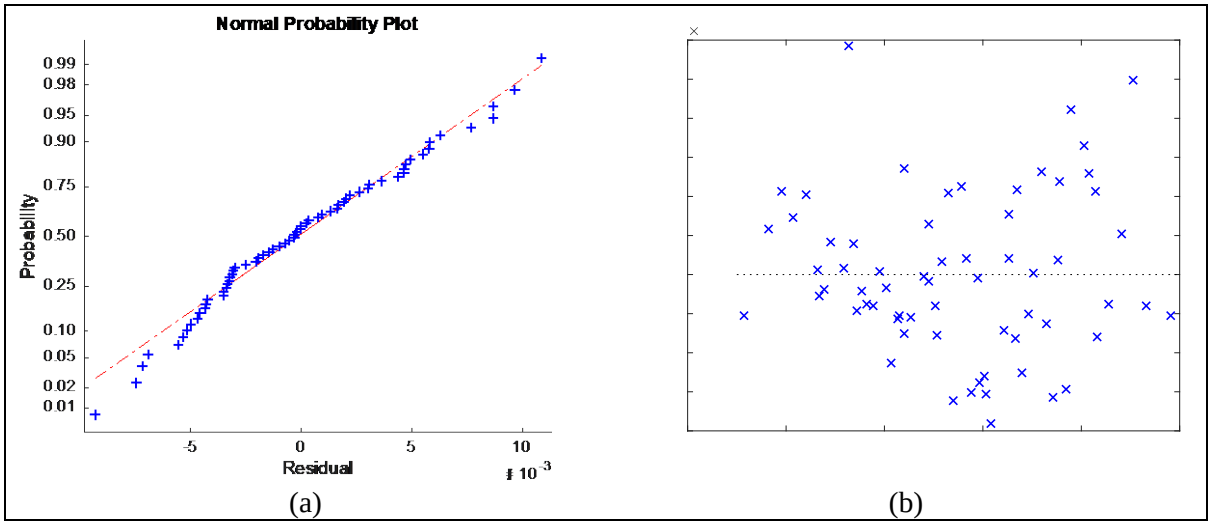
3.1.2.4 Halophia spinulosa

The 3-way ANOVA test for Halophia spinulosa is shown in Table 7.

Table 7- Result of 3-way ANOVA – Halophia spinulosa

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	0.00023	3	7.67E-5	22.4	**0**
Mutation	0.00005	3	1.58E-5	4.61	**0.0099**
Range	0.00005	3	1.69E-5	4.94	**0.0073**
Crossover * Mutation	0.00006	9	6.66E-6	1.95	0.0873
Crossover* Range	0.00004	9	4.68E-6	1.37	0.2506
Mutation*Range	0.00002	9	2.40E-6	0.7	0.7015
Error	0.00009	27	3.42E-6		
Total	0.00054	63			

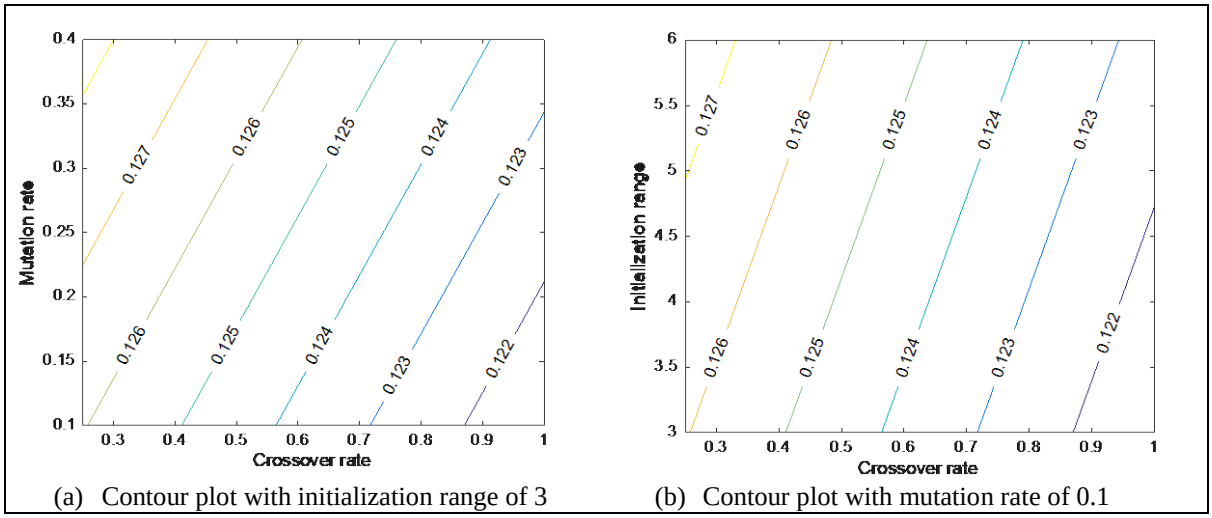
The result of ANOVA shows statistical significance of crossover rate, mutation rate, and initialization range. However, there is no significant interaction between the variables. The check for the validity of ANOVA is presented in Figure 28.



(a) (b)

Figure 28 – (a) Normal probability plot for residuals and (b) residuals versus fitted value – Halophia spinulosa

Both normal probability plot of residuals and plot of residuals versus fitted values in Figure 28. These plots appear to be satisfactory, so there can be ascertained that the conclusion of ANOVA is valid. The response surface that explains the relation of the parameters to the misclassification rate is shown in Figure 29.



(a) Contour plot with initialization range of 3 (b) Contour plot with mutation rate of 0.1

Figure 29 - Contour plot of misclassification rate – Halophia spinulosa

The response surface contour plots shown in Figure 29 imply that the minimization of misclassification rate favours high crossover rate, low mutation rate, and low initialization range.

3.1.2.5 Syringodium isoetofolium

The result of 3-way ANOVA for *Syringodium isoetofolium* is presented in Table 8 below.

Table 8 - Result of 3-way ANOVA – Syringodium isoetofolium

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	0.00072	3	0.00024	18.7	**0**
Mutation	0.00015	3	0.00005	3.89	**0.0196**
Range	0.00182	3	0.00061	47.36	**0**
Crossover * Mutation	0.00008	9	0.00001	0.69	0.7142
Crossover* Range	0.0001	9	0.00001	0.84	0.5883
Mutation*Range	0.00017	9	0.00002	1.49	0.2028
Error	0.00035	27	0.00001		
Total	0.00338	63			

The ANOVA result shows that all three measured parameters i.e. crossover rate, mutation rate and range are statistically significant. However, there are no significant interaction among the parameters. The investigation on the assumptions of ANOVA is shown in Figure 30.

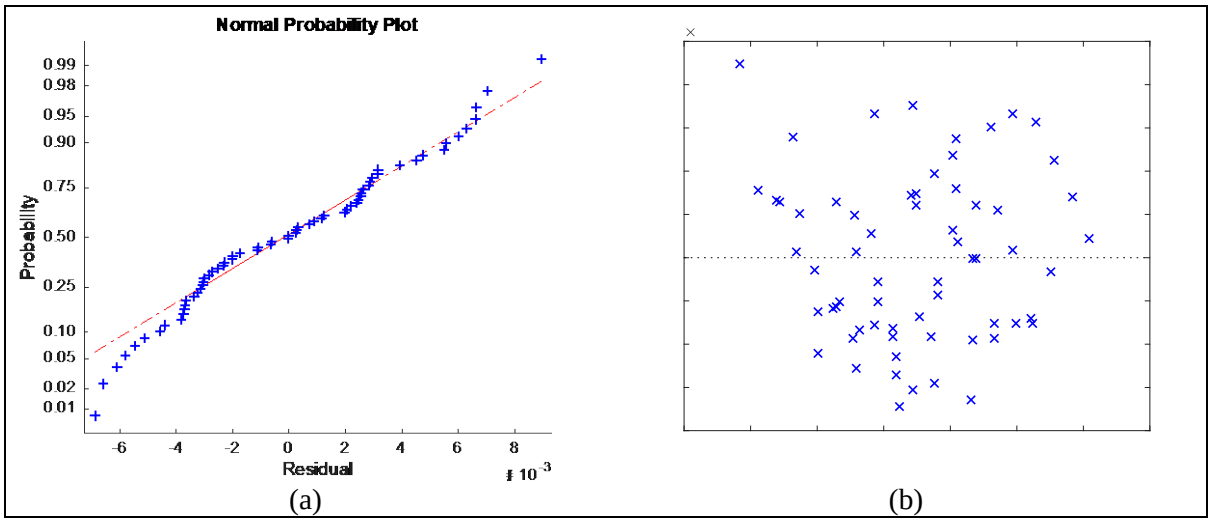
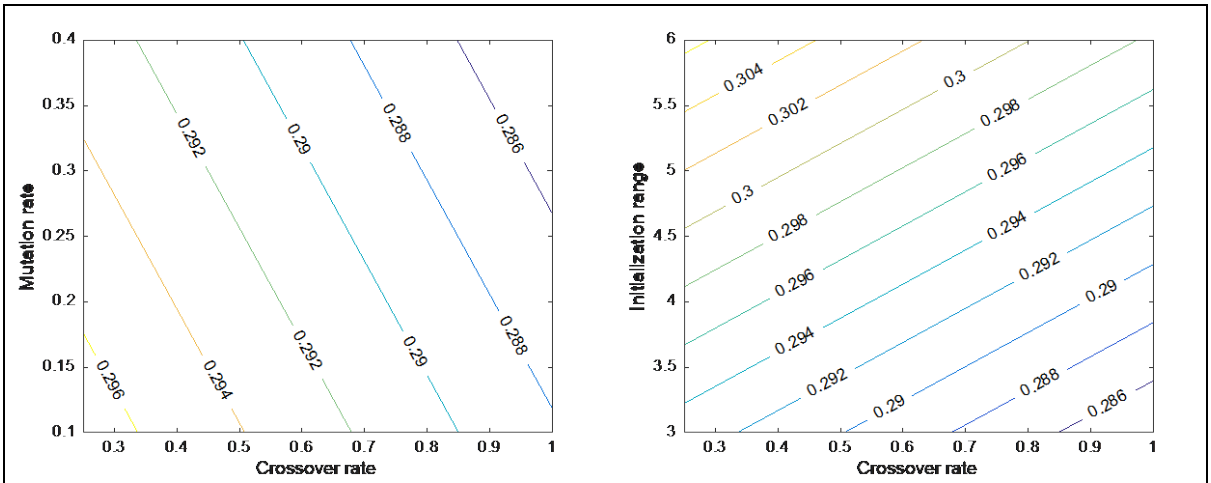


Figure 30- (a) normal probability plot of residuals and (b) residuals versus fitted values – Syringodium isoetofolium

The normal probability plot of residuals in Figure 30a appears to be satisfactory, there is no reason to suspect the normality assumption of ANOVA is invalid. At the first glance, the residuals versus fitted values plot in Figure 30b shows a possibility of autocorrelation. Therefore, a Durbin-Watson test is applied and the result shows that the residuals from linear regression are not correlated at 95% confidence. The ANOVA can then be safely assumed to be valid. The relation of the parameters to the misclassification rate of MOEA is displayed in Figure 31.



(a) Contour plot with initialization range of 3 (b) Contour plot with mutation rate of 0.1

Figure 31 – Contour plot of misclassification rate – Syringodium isoetofolium

Figure 31a shows the response surface contour plot when the initial chromosomes starts with range ($\bar{x} - \sigma$ to $\bar{x} + m\sigma$) where m is 3. Figure 31 is the response surface contour plot when the mutation rate is 0.1. Both contour indicate that minimization of misclassification rate requires crossover rate and mutation rate to be high and initialization range to be low.

3.1.2.6 Zostera muelleri

The result of three-way ANOVA is shown below in Table 9.

Table 9 - Result of 3-way ANOVA – Zostera muelleri

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	0.00025	3	8.48E-5	13.15	**0**
Mutation	0.00001	3	3.21E-6	0.5	0.686
Range	0.00016	3	5.32E-5	8.26	**0.0005**
Crossover * Mutation	0.00008	9	8.69E-6	1.35	0.2598
Crossover* Range	0.00002	9	2.52E-6	0.39	0.9283
Mutation*Range	0.00007	9	7.57E-6	1.17	0.3495
Error	0.00017	27	6.44E-6		
Total	0.00077	63			

The result of ANOVA indicates that crossover rate and initialization range are the statistically significant parameters. The mutation rate appears to have no significance on the misclassification rate as the mutation's fine tuning does not seem to improve or worsen the misclassification rate. To confirm if the conclusion of ANOVA is correct, the normality of the residuals and autocorrelation of the residuals are investigated and the result is shown in Figure 32.

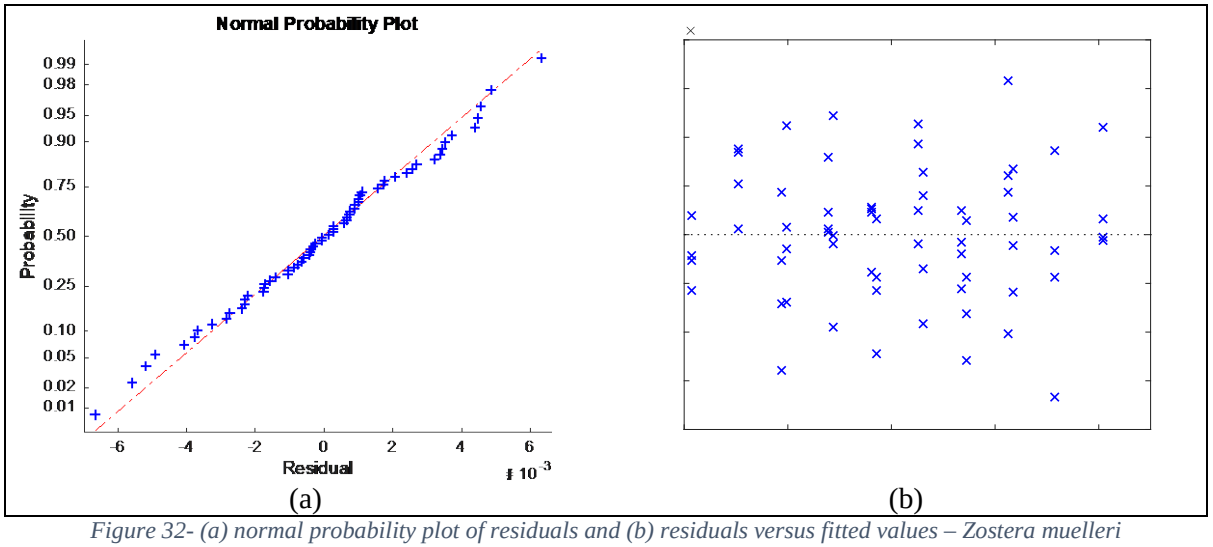


Figure 32- (a) normal probability plot of residuals and (b) residuals versus fitted values – Zostera muelleri

The normal probability of residuals and residuals versus fitted values plots in Figure 32 confirm the valid conclusion of ANOVA. To analyse the relation of the significant parameters, the response surface and contour plot are generated by regression model which are shown in Figure 33. Since the mutation rate is found to be non-significant, therefore the response surface can be plotted with crossover rate and initialization range variables in Figure 33(a).

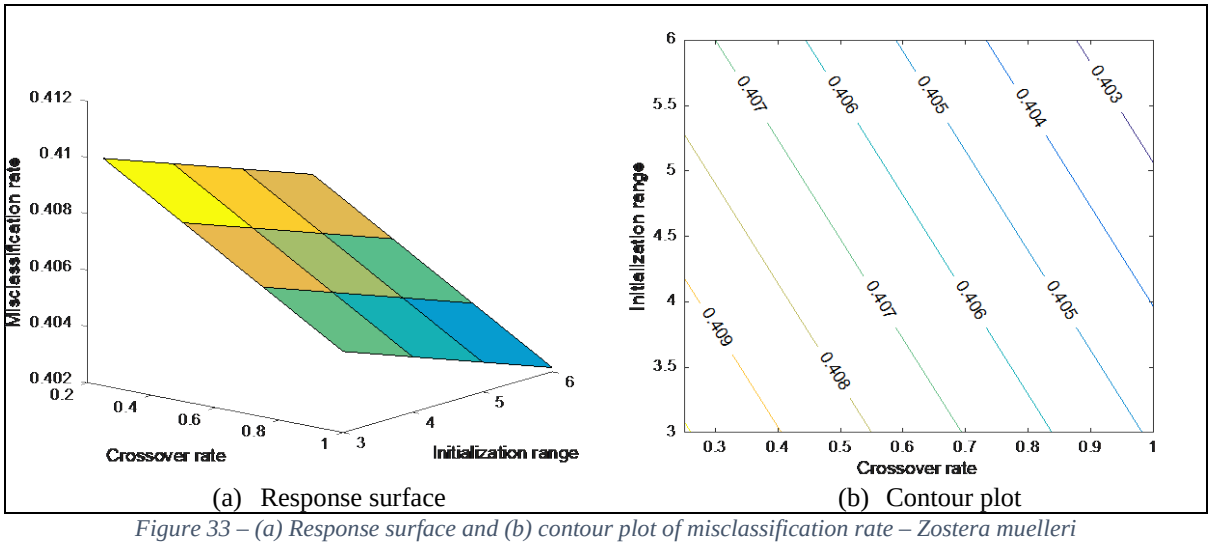


Figure 33 – (a) Response surface and (b) contour plot of misclassification rate – Zostera muelleri

Figure 33 indicates that the low misclassification rate favours high crossover rate and high initialization range.

3.1.2.7 Eucalyptus

The investigation of three-way ANOVA for Eucalyptus is presented in Table 10.

Table 10 - Result of 3-way ANOVA – Eucalyptus

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	0.00233	3	0.00078	64.15	**0**
Mutation	0.00059	3	0.0002	16.21	**0**
Range	0.00016	3	0.00005	4.37	**0.0124**
Crossover * Mutation	0.00012	9	0.00001	1.13	0.3743
Crossover* Range	0.00013	9	0.00001	1.21	0.3275
Mutation*Range	0.00016	9	0.00002	1.49	0.2018
Error	0.00033	27	0.00001		
Total	0.00382	63			

The ANOVA result indicates that the statistical significant correlation of crossover rate, mutation rate, and initialization range towards misclassification rate. However, there is no statistical significance between the parameters. The examination of ANOVA's validity is presented in Figure 34.

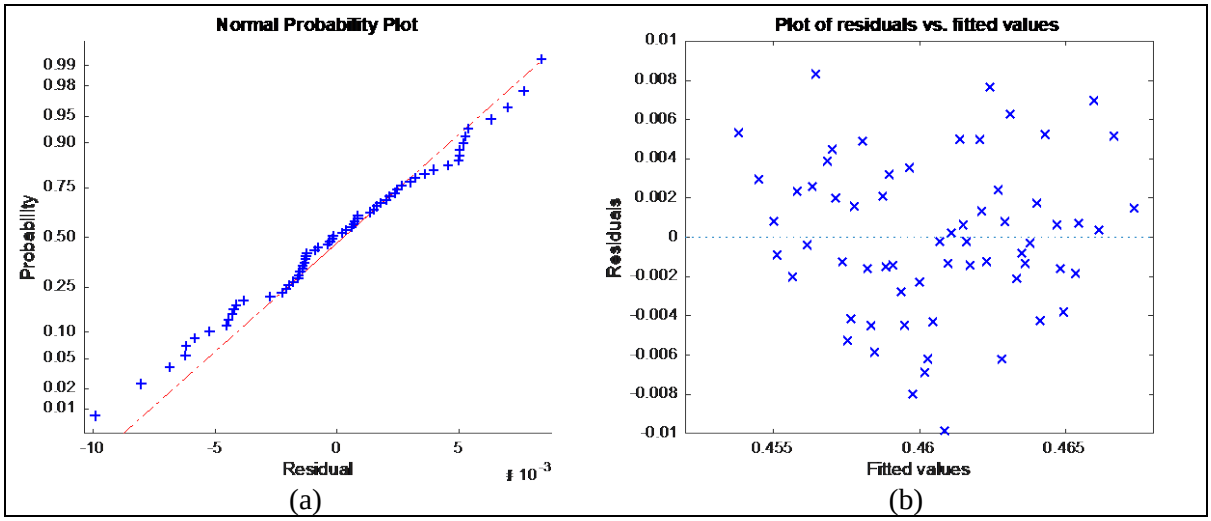
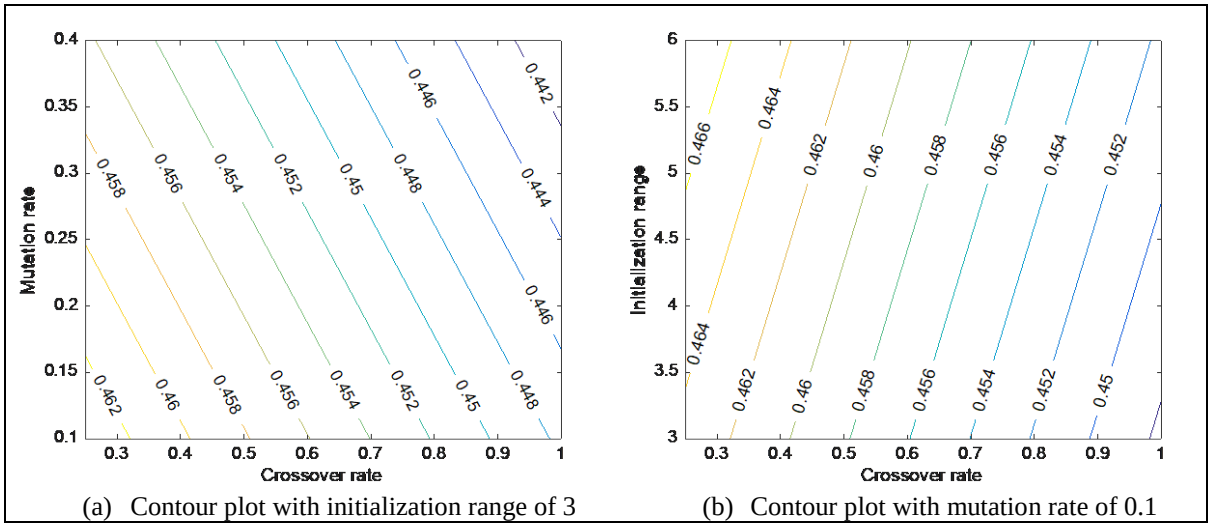


Figure 34- (a) normal probability plot of residuals and (b) residuals versus fitted values – Eucalyptus

Figure 34 presents a normal probability plot of residuals and plot of residuals versus fitted values. These plots appear to be satisfactory, so there is no reason to suspect that there is any problem with the ANOVA's validity. The response surface generated by the regression model is presented in Figure 35.



(a) Contour plot with initialization range of 3 (b) Contour plot with mutation rate of 0.1

Figure 35 – Contour plot of misclassification rate - Eucalyptus

The surface response contour plots in Figure 35(a) and (b) imply that the minimization of misclassification rate requires high crossover rate, high mutation rate, and low initialization range (m = 3).

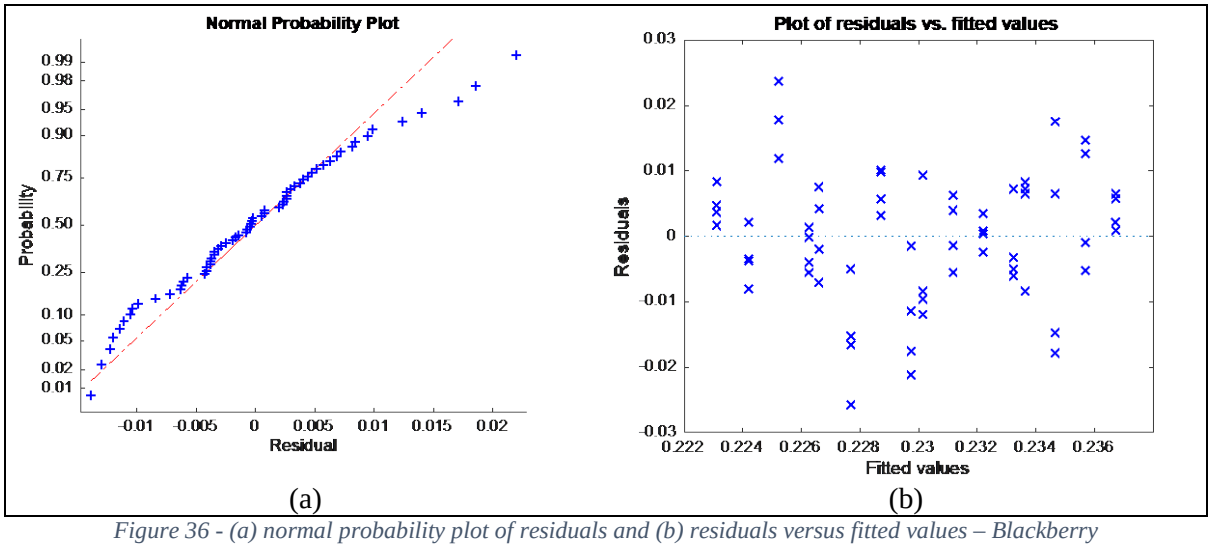
3.1.2.8 Blackberry

The result of three-way ANOVA for blackberry is shown below in Table 11.

Table 11- Result of 3-way ANOVA – Blackberry

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	0.00726	3	0.00242	44.95	**0**
Mutation	0.00076	3	0.00025	4.69	**0.0092**
Range	0.00135	3	0.00045	8.33	**0.0004**
Crossover * Mutation	0.00039	9	0.00004	0.81	0.6152
Crossover* Range	0.00034	9	0.00004	0.71	0.6969
Mutation*Range	0.00079	9	0.00009	1.62	0.1596
Error	0.00145	27	0.00005		
Total	0.01234	63			

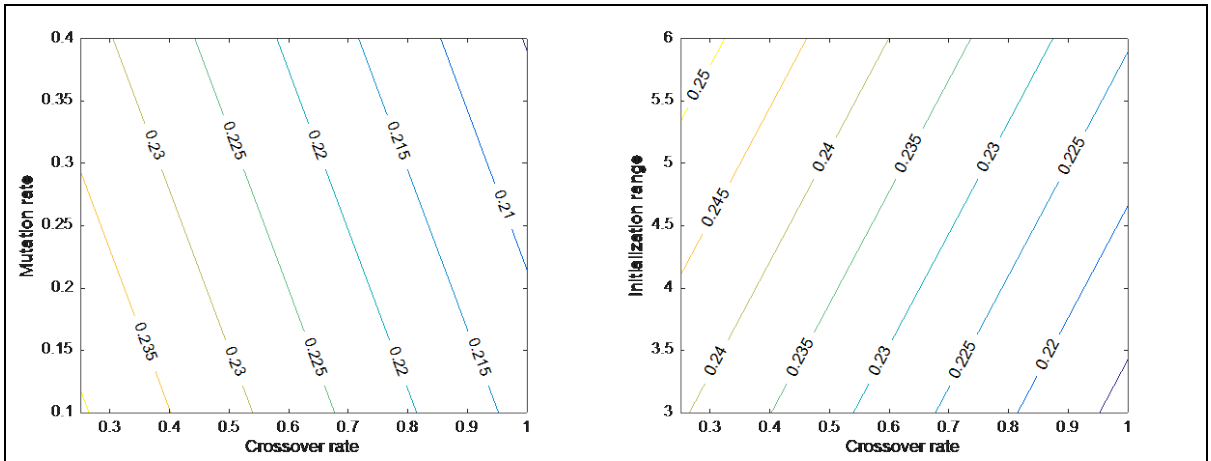
The result shown above in Table 11 imply the statistical significance of crossover rate, mutation rate, and initialization range. To check if ANOVA is valid, the investigations of normality of residuals and non-correlation of the residuals versus fitted value are presented in Figure 36 - (a) normal probability plot of residuals and (b) residuals versus fitted values – Blackberry(a) and (b).



(a) (b)

Figure 36 - (a) normal probability plot of residuals and (b) residuals versus fitted values – Blackberry

Figure 36(a) shows slight skew on the normal probability plot of residuals. A further investigation using Jarque-Bera test shows that the normality distribution is not violated. The residuals versus fitted plot in Figure 36(b) appears to be satisfactory. Therefore, there is no doubt about the validity of ANOVA. The response surface contour plots as displayed in Figure 37 show the relations of the parameters to the misclassification rate.



(a) Contour plot with initialization range of 3 (b) Contour plot with mutation rate of 0.1

Figure 37 - Contour plot of misclassification rate – Blackberry

Figure 37(a) displays the response surface contour plot where the initial chromosome starts from range ($\bar{x} - \sigma$ to $\bar{x} + m\sigma$) where m is 3. Figure 37(b) presents the response surface contour plot when mutation rate is 0.1. These two plots indicate that the low misclassification rate favours high crossover rate, high mutation rate, and low initialization range (m = 3).

3.1.2.9 Lemon (Citrus limon)

The result for three-way ANOVA for lemon is presented below in Table 12.

Table 12- Result of 3-way ANOVA – Lemon

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	5.39E-5	3	1.79E-5	89.27	**0**
Mutation	1.79E-6	3	5.98E-7	2.97	**0.0496**
Range	1.02E-6	3	3.41E-7	1.69	0.1917
Crossover * Mutation	3.90E-6	9	4.33E-7	2.15	0.0601
Crossover* Range	1.46E-6	9	1.63E-7	0.81	0.6115
Mutation*Range	1.80E-6	9	2.01E-7	1.01	0.4654
Error	5.44E-6	27	2.01E-7		
Total	6.94E-5	63			

The ANOVA result shows that only crossover and mutation rate are significant to the misclassification rate. In order to ascertain the assumptions of ANOVA are valid, the normal probability plot of residuals and residuals versus fitted values plot are displayed in Figure 38.

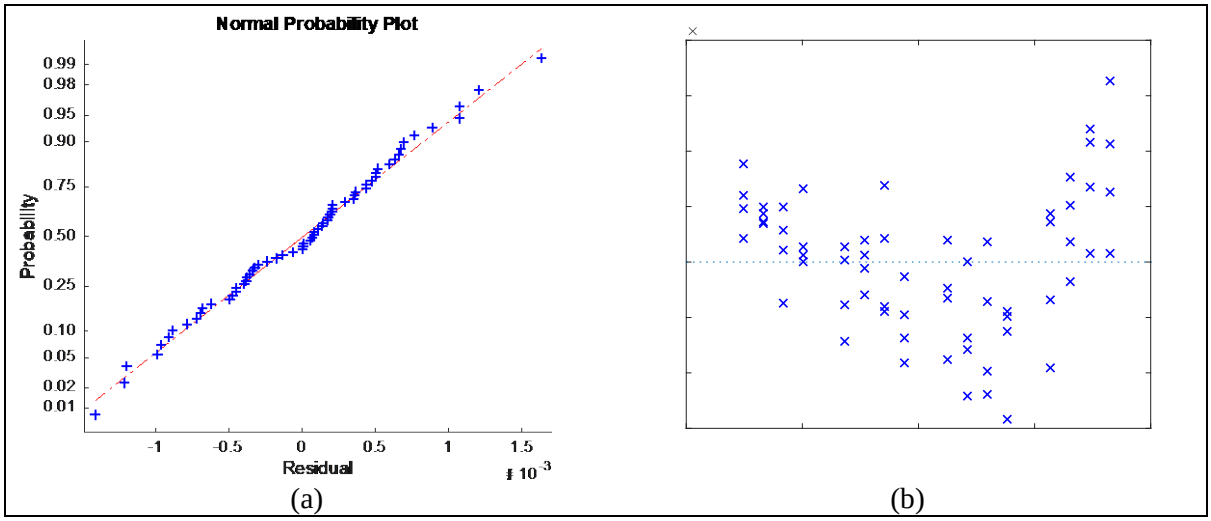
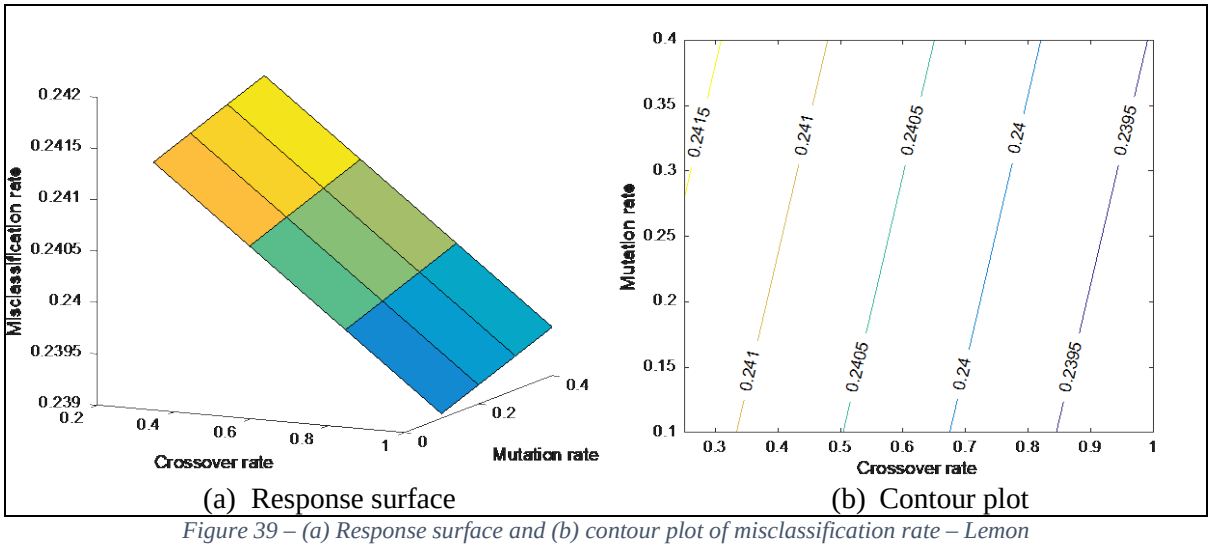


Figure 38 - (a) normal probability plot of residuals and (b) residuals versus fitted values – Lemon

Figure 38(a) presents the normality of the residuals and Figure 38(b) shows that there is no apparent correlation between the variables on the residuals. Therefore, there can be ascertained that the assumptions of ANOVA are valid. The response surface and the contour plot that show the relation of the variables are shown in the Figure 39 below. Since the initialization range is found to be non-significant, the response surface in Figure 39(a) is plotted based on crossover and mutation rate variables.



(a) Response surface (b) Contour plot

Figure 39 – (a) Response surface and (b) contour plot of misclassification rate – Lemon

The surface response and contour plot in Figure 39 suggest that the minimization of misclassification rate favours the high crossover rate and low mutation rate. In this case, the crossover rate is to be 1, and mutation rate to be 0.1.

3.1.2.10 Cherry Plum

The result for a three-way ANOVA test on Cherry plum is shown below in Table 13.

Table 13 - Result of 3-way ANOVA – Cherry plum

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob > F
Crossover	1.40E-5	3	4.69E-6	36.32	**0**
Mutation	2.76E-6	3	9.20E-7	7.1	**0.0011**
Range	1.21E-6	3	4.03E-7	3.12	**0.0426**
Crossover * Mutation	3.13E-6	9	3.48E-7	2.69	**0.0224**
Crossover* Range	2.43E-6	9	2.70E-7	2.09	0.0678
Mutation*Range	1.86E-6	9	2.07E-7	1.6	0.165
Error	3.49E-6	27	1.29E-7		
Total	2.90E-5	63			

The result of ANOVA implies that crossover rate, mutation rate, and initialization range are statistically significant. In addition, the interaction of crossover rate and mutation rate is also found to be significant. To check if the conclusion is valid, the investigation on the normality of the residuals and non-correlation of variables on residuals is shown in Figure 40.

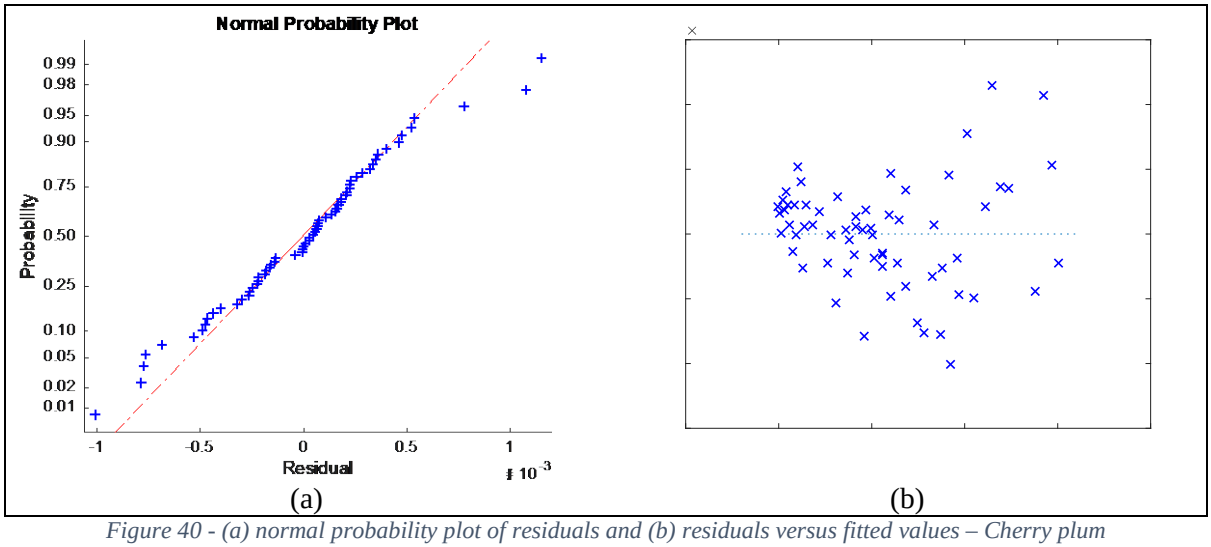


Figure 40 - (a) normal probability plot of residuals and (b) residuals versus fitted values – Cherry plum

The normal probability plot of residual in Figure 40(a) shows that the result is slightly skewed. A Jarque –Bera test is applied and shown that the normality distribution is still acceptable. The Dawson-Durbin test is also applied on the residuals versus the fitted value (Figure 40b) which supports that there is non-correlation among variables on the residuals. The response surface is then generated from the regression model and is shown in Figure 41.

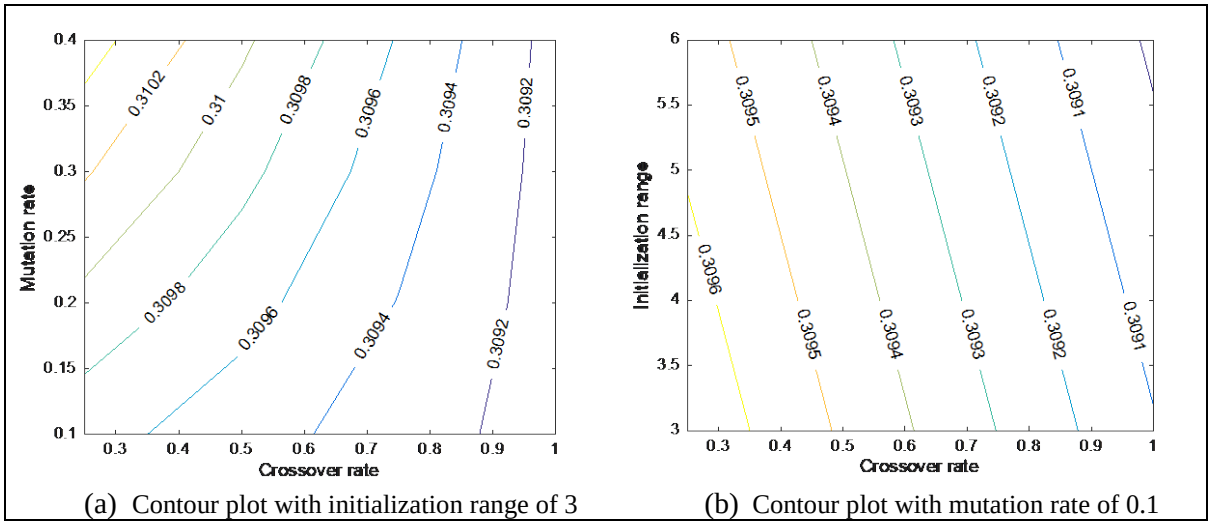


Figure 41 - Contour plot of misclassification rate – Cherry plum

The response surface contour plots in Figure 41 implies that the minimization of misclassification rate favours the high crossover rate, low mutation rate, and high initialization range.

3.2 Summary of Optimum Parameters

For the optimum parameters, the high crossover rate is fixed as 1, low mutation rate as 0.1, high mutation rate as 0.4, low initialization range as 3, and high initialization range as 6. The tabulated result of the optimum parameters is presented below in Table 14.

Table 14 - List of the optimum parameters

Plant species	Optimum crossover rate	Optimum mutation rate	Optimum Initialization range	Optimum number of generation
Cymadocea serrulata	1	0.4	3	300
Halodue univervis	1	0.4	6	230
Halophia ovalis	1	0.1	3	310
Halophia spinulosa	1	0.1	3	590
Syringoidum isoetofolium	1	0.4	3	360
Zostera muelleri	1	N/A*	6	270
Eucalyptus	1	0.4	3	190
Blackberry	1	0.4	3	300
Lemon	1	0.1	N/A*	170
Cherry plum	1	0.1	6	120

** N/A when the particular variable is found to be insignificant to the misclassification rate*

As expected, the optimum crossover rate needs to be high so the MOEA can search for the optimum solution faster. The optimum mutation rate surprisingly varies between 0.1 and 0.4. However, this can be attributed that higher mutation rate allows the MOEA to fine tune the solution faster hence it favours higher mutation rate. Lower mutation rate is also expected since mutation rate can also deteriorate solutions. Therefore, it depends on case by case basis, whether the MOEA prefers high or low mutation rate.

The high or low initialization range depends on the location of global optima. If the location of global optima is closer to the mean, then low initialization range is preferred and vice versa.

3.3 Calculated Threshold Values from Unoptimized MOEA, Optimized MOEA, and Adaptive Threshold Method

Table 15 presents the calculated threshold values using the unoptimized MOEA, optimized MOEA and adaptive threshold method. The calculation of the adaptive threshold was slightly modified to be mean

plus three times standard deviation as the used measures for every spectral similarity measure(Schwarz and Staenz, 2001).

Table 15- The threshold value for MOEA, optimized MOEA and adaptive threshold

		Unoptimized MOEA threshold	Optimized MOEA threshold	Adaptive Threshold
Cymadocea Serrulata	SAM	0.0819	0.0648	0.228
	SCA	0.1057	0.0387	0.2858
	SID	0.0292	0.0125	0.041
	SID-SAM	0.0021	0.0007	0.0055
	nSSM	0.167	0.3229	0.4428
Halodue univervis	SAM	0.1061	0.0703	0.2426
	SCA	0.1245	0.0597	0.2506
	SID	0.0192	0.0063	0.0762
	SID-SAM	0.0016	0.0004	0.0116
	nSSM	0.1813	0.1182	0.2912
Halophia ovalis	SAM	0.0808	0.0807	0.2369
	SCA	0.0964	0.0735	0.3092
	SID	0.0222	0.0162	0.0403
	SID-SAM	0.0026	0.0016	0.006
	nSSM	0.2047	0.1338	0.3152
Halophia spinulosa	SAM	0.0754	0.0592	0.146
	SCA	0.0879	0.0344	0.1712
	SID	0.0148	0.0075	0.0235
	SID-SAM	0.0014	0.0003	0.0021
	nSSM	0.1083	0.1519	0.5662
Syringodium isoetifolium	SAM	0.1332	0.0571	0.1901
	SCA	0.1237	0.0526	0.1998
	SID	0.0146	0.0087	0.0422
	SID-SAM	0.0027	0.0006	0.0047
	nSSM	0.1709	0.3787	0.8617
Zostera muelleri	SAM	0.0911	0.0832	0.2224
	SCA	0.14	0.0612	0.2197
	SID	0.0123	0.0166	0.0718
	SID-SAM	0.0012	0.0013	0.0099
	nSSM	0.2497	0.1330	0.5908
Eucalyptus *(Eucalyptus globulus)*	SAM	0.0874	0.0751	0.2621
	SCA	0.1120	0.0942	0.2002
	SID	0.0292	0.0163	0.2038
	SID-SAM	0.0035	0.0020	0.0401
	nSSM	0.1341	0.1625	0.6584
Blackberry *(Rubus fruticosus agg)*	SAM	0.0868	0.0723	0.3245
	SCA	0.0747	0.0599	0.2647
	SID	0.0455	0.0411	0.1701
	SID-SAM	0.0058	0.0034	0.0415
	nSSM	0.1764	0.1235	0.5308
Lemon *(Citrus limon)*	SAM	0.1003	0.0796	0.4066
	SCA	0.4424	0.0646	0.3006
	SID	0.0713	0.0277	0.3632
	SID-SAM	0.0078	0.1776	0.1212
	nSSM	0.1124	0.1266	0.6734
Cherry Plum *(Prunus cerasifera)*	SAM	0.0837	0.0737	0.3402
	SCA	0.0732	0.0664	0.1863
	SID	0.0193	0.0220	0.2811

	SID-SAM	0.0011	0.0009	0.1102
	nSSM	0.1832	0.2127	0.5673

The performance of the MOEA in terms of the rate of misclassification is shown in Figure 42. For comparison purposes, the performance of adaptive threshold method and Support Vector Machine (SVM) algorithm is also presented in the same figure. The multi-class SVM implementation in MATLAB's Statistics and Machine Learning Toolbox is used in this thesis.

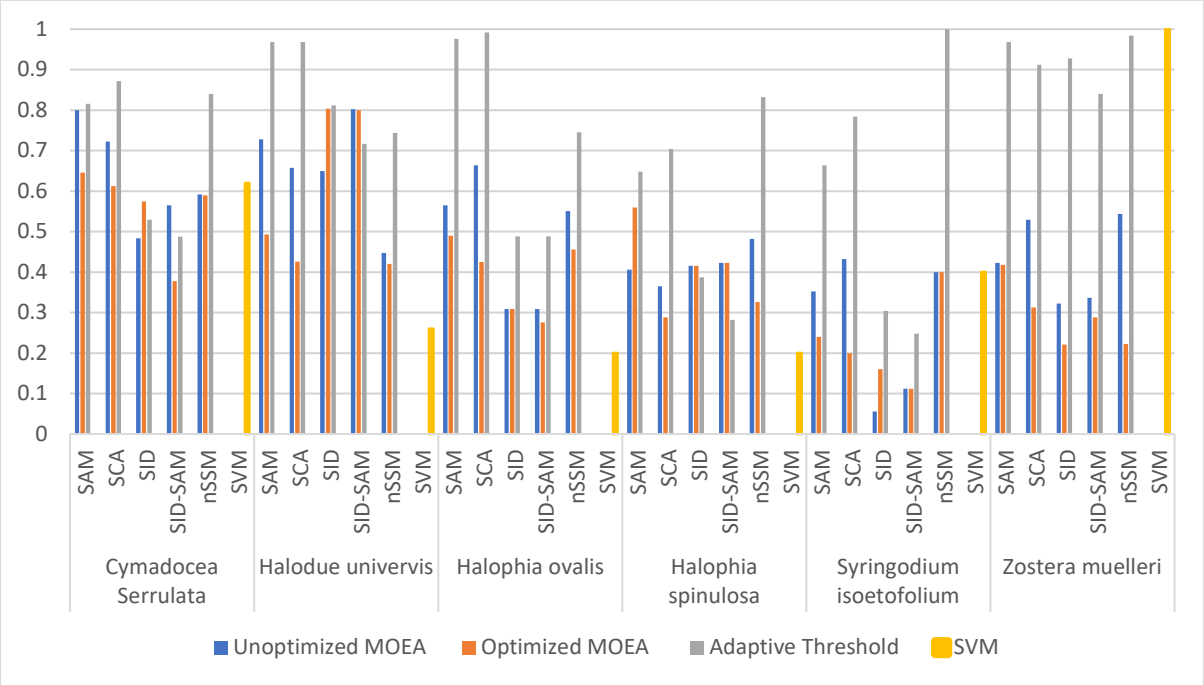


Figure 42 - Comparison of misclassification rate among spectral similarity measures for seagrass

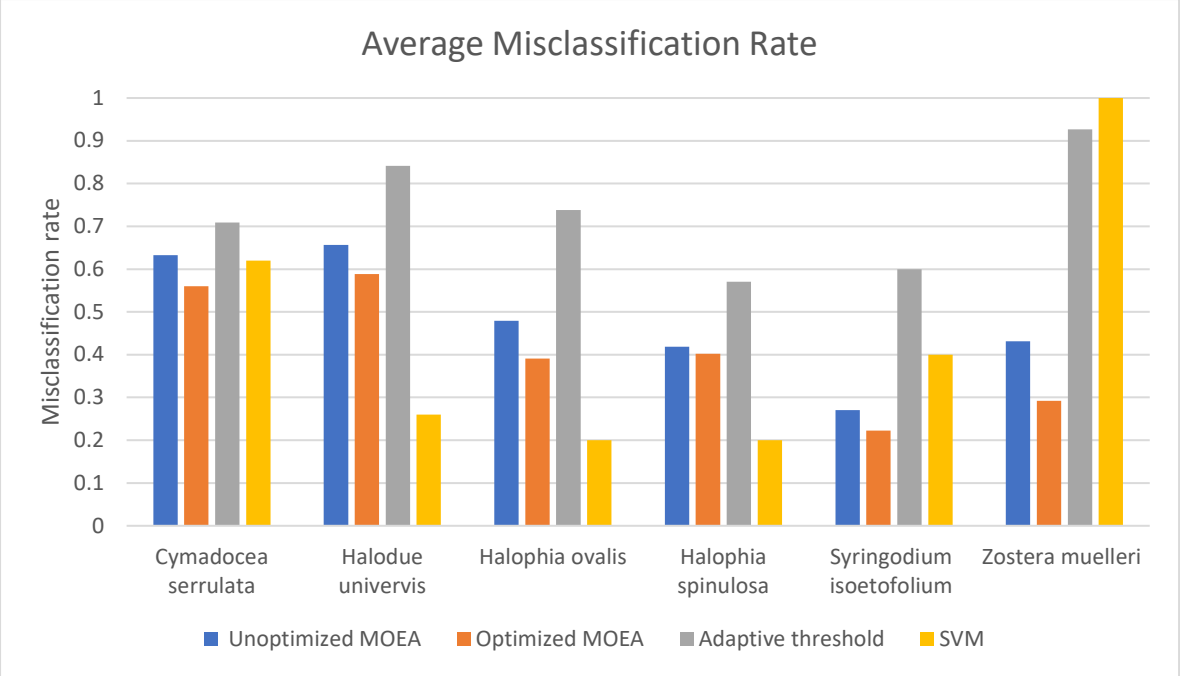


Figure 43 - Average misclassification rate among seagrasses for different methods

The comparison of the misclassification rate for seagrasses shown in Figure 42 and Figure 43 indicates that:

1. The performance of optimized MOEA was more superior than unoptimized MOEA in 70% of the analysis cases;
2. In 5 out of 30 analysis cases, the difference in misclassification rate between unoptimized and optimized MOEA is less than 1%;
3. Both unoptimized and optimized MOEA perform better than adaptive threshold method in 83.3% of the analysis cases;
4. SID and SID-SAM give lower misclassification rates than other similarity measures in 50% of the analysis cases;
5. SVM performs better than SID and SID-SAM 50% of the analysis cases
6. MOEA has lowest misclassification rate for *Syringodium isoetifolium*, while SVM performs best for *Halophia ovalis* and *Halophia spinulosa*;
7. MOEA has highest misclassification rate for *Halodue univervis*, while SVM performs worst for *Zostera muelleri*; and

8. *Halophia ovalis* and *Halophia spinulosa* have almost identical misclassification rate. This is expected since both have very close taxonomical relationship, i.e. belonging to the same genus.

Comparison of the misclassification rate among the different spectral similarity measures for terrestrial plants shown in Figure 44 and Figure 45. The following observations can be made:

1. The performance of optimized MOEA is better than unoptimized MOEA in 80% of the analysis cases;
2. The optimized MOEA performed relatively equal with the unoptimized MOEA 5% of the analysis cases;
3. The unoptimized MOEA performed better than optimized MOEA in 15% of the analysis cases;
4. The adaptive threshold method performed worse than unoptimized MOEA and optimized MOEA all the time;
5. SVM performed better than unoptimized MOEA and optimized MOEA in 75% of the analysis cases;
6. The optimization on MOEA sometimes worsen nSSM and SID-SAM misclassification rate; and
7. Blackberry (*Rubus fruticosus agg.*) has the lowest misclassification rate for all the methods.

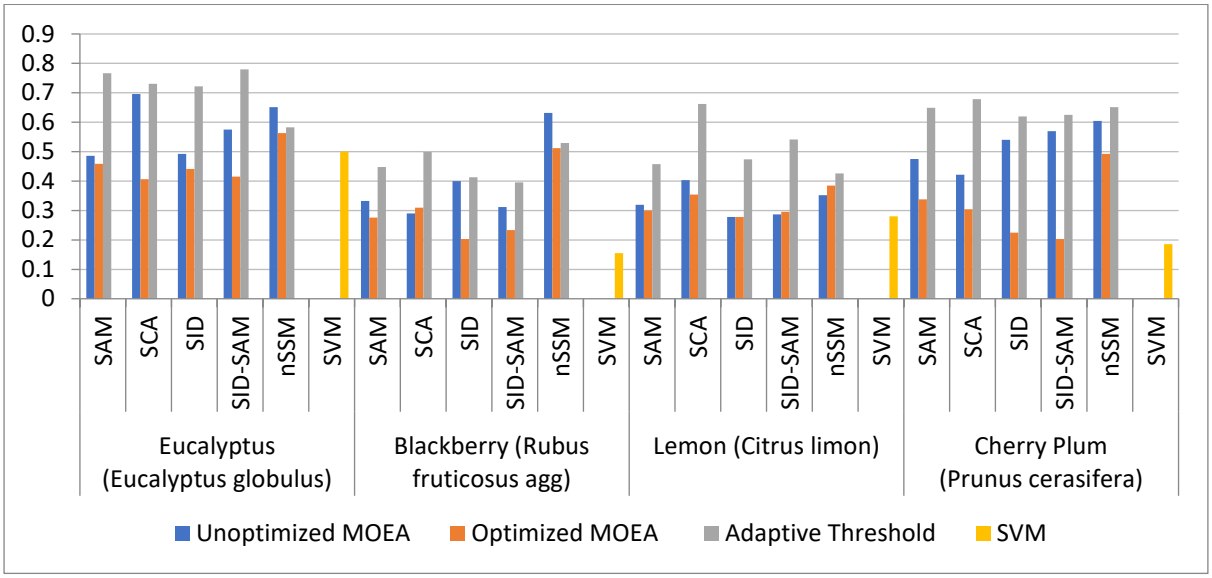


Figure 44 - Comparison of misclassification rate among spectral similarity measures for terrestrial plants

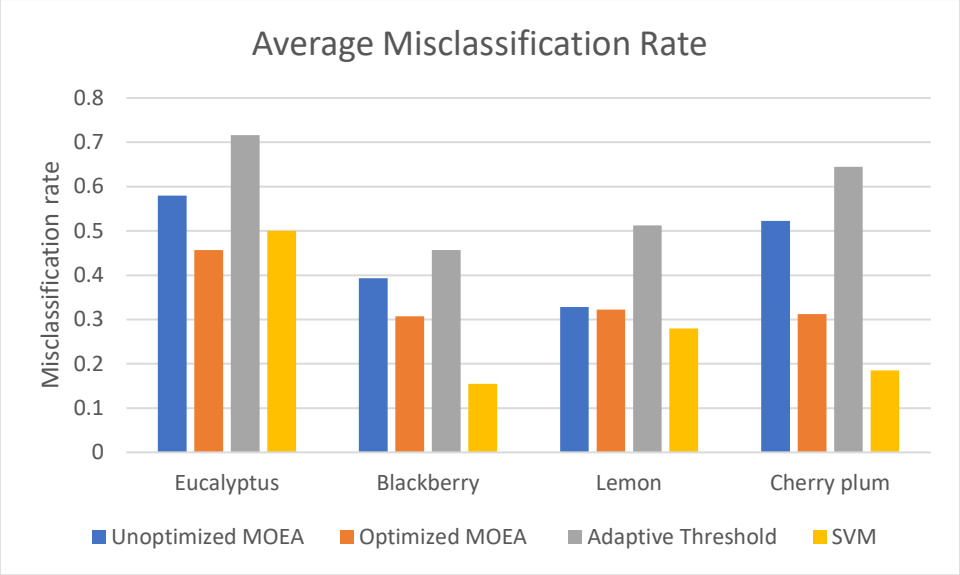


Figure 45 - Average misclassification rate among terrestrial plants for different methods

3.4 Effectiveness of Different Spectral Similarity Measures in Classifying One Species

Different spectral similarity measures have different effectiveness in classifying one species from a given set of spectral signatures obtained from different samples of that particular species. The effectiveness in this regard is quantified by the Probability of Spectral Discrimination (PSD) (Chang, 2003). The PSD calculates the relative probability to distinguish a target spectral signature from reference signatures in a library or database. Let $\{s_l\}_{l=1}^{L}$ be L spectral signatures from a library / database Δ and t is the target spectral signature to be identified from Δ. The PSD of all s_k in database Δ with respect of t defined as:

$$p_{t,\Delta}^{m}(l) = \frac{m(t, s_l)}{\sum_{i=1}^{L} m(t, s_i)} \quad \text{for } l = 1, 2, \ldots, L \tag{25}$$

where $m(t, s_l)$ is the sum of a spectral similarity measure of target spectra t compared with all other spectral signatures S_l in the database Δ, and $\sum_{i=1}^{L} m(t, s_i)$ is the normalization constant determined by t and Δ. The result $p_{t,\Delta}^{m} = (p_{t,\Delta}^{m}(1), p_{t,\Delta}^{m}(2), \ldots, p_{t,\Delta}^{m}(L))^{T}$ is the magnitude of PSD of Δ with respect to t.

Higher given PSD value of one spectral measures indicates that particular measure is more amenable to differentiate spectral signature from a set of spectra. In general, the endmember can be more accurately discriminated against the other endmember signatures. The result of PSD is shown in Figure 46 for different types of seagrass and Figure 47 for four types of terrestrial plants.

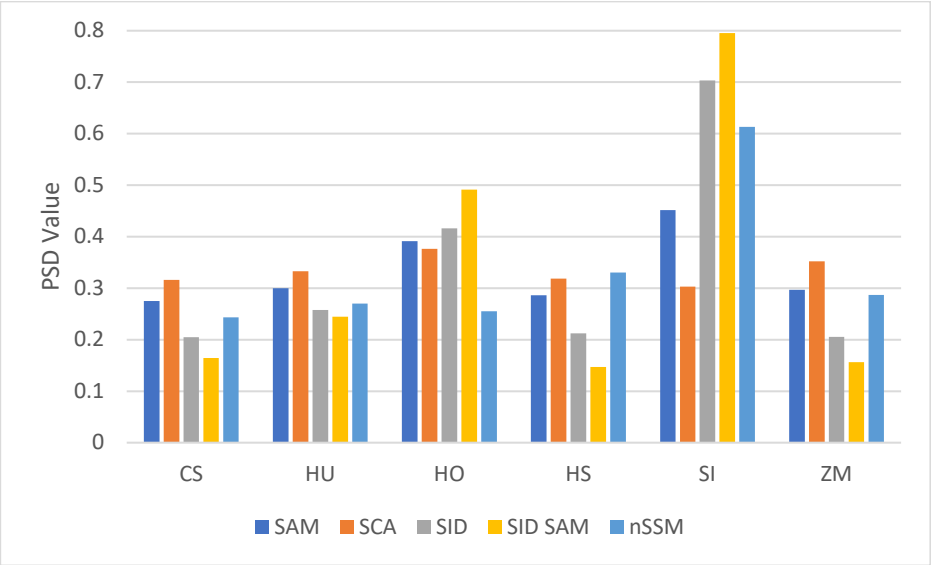


Figure 46 - PSD for different spectral measures applied to different type of seagrass

In Figure 46 *Syringodium isoetofolium* has the highest value of PSD from most of the similarity measures; this means its spectra are the most distinctive. Therefore, it can be identified easily using four out of five similarity measures i.e. SAM, SID, SID-SAM and nSSM. In terms of the performance of the measures stochastic and hybrid measures SID-SAM and SID, perform better than the other measures on identifying *Syringodium isoetofolium* and *Halophia ovalis*. However, they are very ineffective in identifying other species of seagrasses. SCA consistently performs better than SID and SID-SAM for the rest of the seagrass except *Halophia spinulosa* because it takes into account the overall shape of reflectance signal.

The PSD for terrestrial plants is shown in Figure 47. It shows that blackberry has the highest PSD value, which implies that its spectra are the most distinctive from those of other species. In contrast to their application to seagrass, SID and SID-SAM are less capable in discriminating terrestrial plants. They are outperformed by SCA, and nSSM in 70% of the cases.

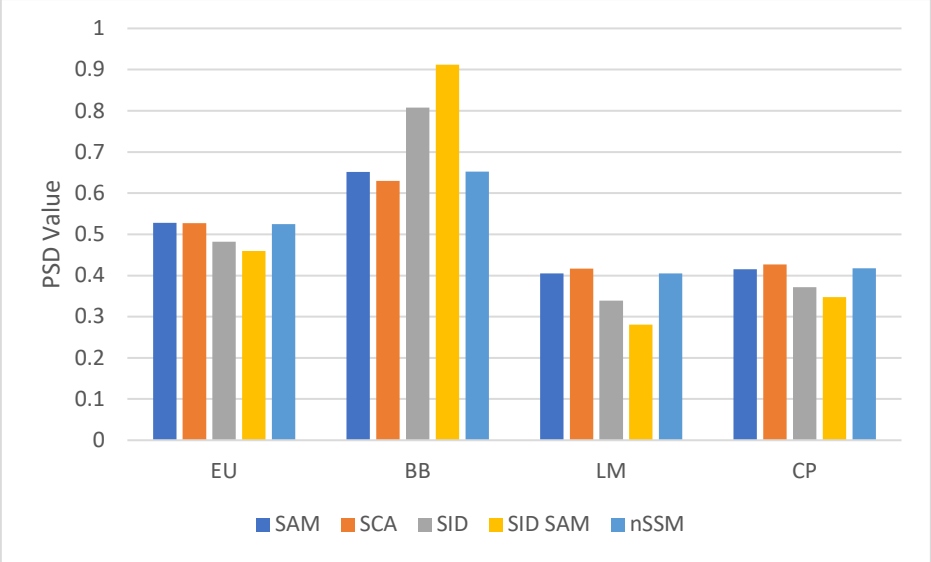


Figure 47 – PSD statistics for different similarity measures on terrestrial plants

3.5 Effectiveness of Different Spectral Similarity Measures in Discriminating Spectral Data

Another way of measuring the discriminating capacity of one spectral similarity measure is the Power of Spectral Discrimination (PWSD), which calculates how well one spectral signature can be differentiated from another relative to a reference spectra (Chang, 2000). Let $m(r_i,d)$ and $m(s_i,d)$ be the value of given by one spectral measure which d is the reference spectra and spectral signatures, r_i and s_i . The PWSD selects the maximum of two ratios, ratio of $m(r_i,d)$ to $m(s_i,d)$ and ratio of $m(s_i,d)$ to $m(r_i,d)$, which is shown as follows:

$$\text{PWSD}^m(r_i, s_i; d) = max\left\{\frac{m(r_i, d)}{m(s_i, d)}, \frac{m(s_i, d)}{m(r_i, d)}\right\}. \quad (26)$$

Higher value of PWSD indicates higher discriminating capacity of that particular spectral similarity measure. In most cases PWSD is larger than unity except for $r_i = s_i$ then PWSD is 1. PWSD of five spectral similarity measures applied to seagrass and terrestrial plants is shown in Figure 48, and Figure 49, respectively.

Figure 48 –PWSD of different spectral similarity measures applied to (a) Cymadocea serrulata, (b) Halodue univervis, (c) Halophia ovalis, (d) Halophia spinulosa, (e) Syringodium isoetofolium, and (f) Zostera muelleri

The results of PWSD in Figure 48 support that the indication of PSD values of *Syringodium isoetofolium* and *Halophia ovalis* have the most distinctive spectral signatures. It is also confirmed that SID-SAM has better capacity than the other four measures to discriminate seagrass. This is in line with the finding in (Du et al., 2004).

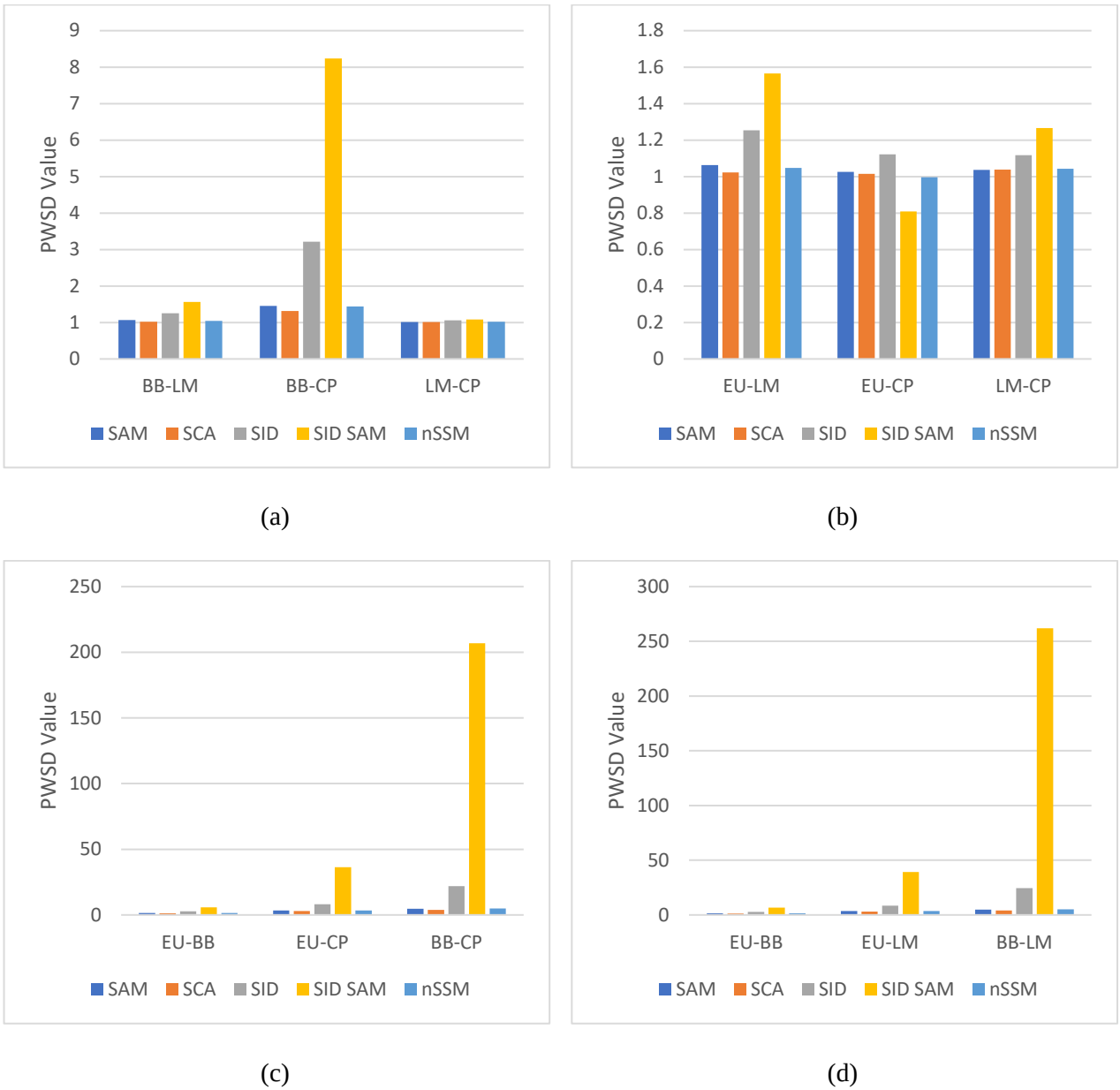


Figure 49- The result of PWSD on (a) Eucalyptus (Eucalyptus globulus), (b) Blackberry (Rubus fruticosus agg.), (c) Lamon (Citrus limon), (d) Cherry Plum (Prunus cerasifera)

The PWSD in Figure 49 suggests that Blackberry has indeed the most distinctive spectra. As in the case for seagrass, the SID-SAM exhibits the highest discriminating capacity.

Chapter 4

Conclusions and Future Research

"There is no deduction that is superior or inferior…because there is only one truth."

Shinichi Kudo

4.1 Discussion

Application of PSD and PWSD to seagrasses confirms that *Syringodium isoetofolium* has the most distinctive spectral signature, which leads to it having the lowest misclassification rate among other species of seagrass. *Halophia ovalis* comes in the second place. Classification of these two particular species can be ascertained by the use of only two spectral similarity measures, namely SID-SAM and SID.

Other species of seagrass are more challenging to classify and require the use of at least three similarity measures to ascertain their classification. The following combinations offer the best likelihood of minimizing misclassification:

1. *Halophia spinulosa*: SCA, nSSM, and SID
2. *Zostera muelleri*: nSSM, SID, and SID-SAM
3. *Cymadocea serrulata*: SID-SAM, SID, and nSSM
4. *Halodue univervis*: nSSM, SID-SAM, and SCA

In the case of terrestrial plants, PSD and PWSD affirms that Blackberry has the most distinctive spectra signature. Classification of different terrestrial plant may be ascertained by the following combinations in the usage of spectral similarity measures:

1. Blackberry: SID-SAM and SID
2. Lemon: SID, SID-SAM, and SAM
3. Eucalyptus: SAM, nSSM, and SCA

4. Cherry plum: SAM, SCA, and nSSM

SVM is found to be as good as than MOEA in classifying seagrass species, and it outperforms MOEA in 75% of the analysis case of terrestrial plants. These results are attributed to the different spectrometers' specification used in the measurement. The ASD-VNIR spectrometer which is used to measure the seagrass has approximate optical resolution of 7 nm from 400 to 1050 nm, whereas Ocean Optics STS-VIS spectrometer which is used to measure terrestrial plants has optical resolution of 1.5 nm from 350-800 nm. Therefore, MOEA results for seagrass species are on-par with SVM as the ASD spectrometer range includes large NIR region. On the other hand, the SVM outperforms MOEA in 75% of the analysis case of terrestrial plants since Ocean Optics spectrometer has lower NIR range and higher resolution than ASD spectrometer. Higher resolution helps SVM to perform better than MOEA as SVM could captures finer shape variation.

4.2 Conclusions

As expected for the outcome, the performance of MOEA has showed lower misclassification rate than the adaptive threshold method in identifying target seagrasses and terrestrial plants. Therefore, it can be concluded that MOEA is more superior to the adaptive threshold method. The SVM which is used as the benchmark comparison for threshold method performed relatively on-par with optimized MOEA and better than adaptive threshold method.

The usage of MOEA should also work on larger and more diverse dataset as the method works on closely related seagrass data, i.e. *Halophia ovalis* and *Halophia spinulosa.* This is further corroborated by the successful application of MOEA on classifying the terrestrial plants.

This research project also investigated the optimization of MOEA by a series of parametric studies on different parameters, i.e. initialization chromosome, number of generation, crossover rate and mutation rate. The outcome greatly improves the effectiveness of MOEA.

Comparison of similarity measures' using PSD and PWSD shows that some measures work better than the other measures for a specific seagrass. Using this information and MOEA, suggestions have been

made as to the preferred usage of several measures for specific seagrass. This will aid a user to make a decision if there is conflicting result from different measures.

4.3 Recommendations for Future Research

Further work to improve the status of the current research includes:

1. Building Spectral Library

The research presented in this thesis is based on ten spectra signatures of six different seagrass species, and 100 spectral signatures of four terrestrial species. Extension to collect spectral signatures of other species is required to facilitate automated classification of all types of plants.

2. In-situ Underwater Measurement

All spectral signatures used in the current research were collected in a controlled environment. Further research on classification based on spectra from in-situ underwater measurement is required in the context of automated classification survey. Complex adaptations are especially needed when the reflectance spectra are measured underwater as they are affected by canopy effect, i.e. the difference in light distribution due to shadow. During the measurement, there can be wave and current which distorts the reflectance spectra measurement. Additionally, the reflection from the sea floor can also affect the measurement if the floor is highly reflective such as sand. Other than that, water molecules attenuate the travel of light especially red colour, therefore the reflectance colour needs to be restored (Bongiorno et al., 2013). There is also limitation on the equipment, particularly for spectrometers which measure light reflectance at one big pixel, and there can be more than one material presents in the measurement. Spectral decomposition (unmixing) needs to be applied before the measured reflectance is being analysed in this case.

3. Enhancement of MOEA

The current research has demonstrated a promising application of the MOEA in the determination of threshold for different similarity measures. Investigation on other variants of MOEA such as VEGA, and SPEA2 ensures robustness of the recommended threshold value.

Recently, there is a new method introduced for hyperspectral classification, i.e. deep learning (Chen et al., 2014). This method is very robust especially for the feature extraction. In the future, the research focuses on combining deep learning for feature extraction of hyperspectral signal and MOEA for threshold classification should be very promising.

References

AASEN, H., BENDIG, J., BOLTEN, A., BENNERTZ, S., WILLKOMM, M. & BARETH, G. 2014. Introduction and preliminary results of a calibration for full-frame hyperspectral cameras to monitor agricultural crops with UAVs. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* 40**,** 1.

AGRAWAL, R. B. 1995. Simulated binary crossover for continuous search space. *Complex systems,* 9**,** 115-148.

BAJWA, S., BAJCSY, P., GROVES, P. & TIAN, L. 2004. Hyperspectral image data mining for band selection in agricultural applications. *Transactions-american Society of Agricultural Engineers,* 47**,** 895-908.

BEYER, H.-G. & DEB, K. On the desired behaviors of self-adaptive evolutionary algorithms. *International Conference on Parallel Problem Solving from Nature,* 2000. Springer, 59-68.

BOISSET, F., 2011, Syringodium isoetifolium (Ascherson) Dandy, Digital Image, AlgaeBase: World-wide electronic publication, National University of Ireland, Galway. Available: http://www.algaebase.org/search/species/detail/?species_id=21542 [Accessed 1 Feb 2018].

BOISSET, F., 2015, Halophia Ovalis(R.Brown), Digital Image, AlgaeBase. Available: http://www.algaebase.org/search/species/detail/?species_id=o6dbf8885486b4260 [Accessed 1 Feb 2018].

BONGIORNO, D. L. 2014. *Hyperspectral benthic mapping from underwater robotic platforms.* PhD Doctorate, University of Sydney.

BONGIORNO, D. L., BRYSON, M. & WILLIAMS, S. B. Dynamic spectral-based underwater colour correction. *OCEANS - Bergen, 2013 MTS/IEEE,* 10-14 June 2013 2013. 1-9.

BOTHA, E. J., BRANDO, V. E., ANSTEE, J. M., DEKKER, A. G. & SAGAR, S. 2013. Increased spectral resolution enhances coral detection under varying water conditions. *Remote Sensing of Environment,* 131**,** 247-261.

BRYCE, C., 2017, Halophia Spinulosa, Digital Image, Western Australia Museum. Available: http://museum.wa.gov.au/explore/marine-life-dampier-archipelago/explore-marine-life/flora [Accessed 1 Feb 2018].

CHANG, C.-I. 2000. An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis. *IEEE Transactions on information theory,* 46**,** 1927-1932.

CHANG, C.-I. 2003. *Hyperspectral imaging: techniques for spectral detection and classification,* Springer Science & Business Media.

CHEN, Y., LIN, Z., ZHAO, X., WANG, G. & GU, Y. 2014. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected topics in applied earth observations and remote sensing,* 7**,** 2094-2107.

CZARN, A. S. T. 2008. *Statistical exploratory analysis of genetic algorithms.* Doctoral Thesis, University of Western Australia.

DAVIE, A., HARTMANN, K., TIMMS, G., DE GROOT, M. & MCCULLOCH, J. Benthic habitat mapping with autonomous underwater vehicles. *OCEANS 2008,* 15-18 Sept. 2008 2008. 1-9.

DAVIS, L. 1991. *Handbook of genetic algorithms,* New York : Van Nostrand Reinhold, c1991.

DEB, K. 2001. *Multi-objective optimization using evolutionary algorithms,* Chichester ; New York : John Wiley & Sons, c2001.

DEB, K., AGRAWAL, S., PRATAP, A. & MEYARIVAN, T. 2000. A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimization: NSGA-II. *In:* SCHOENAUER, M., DEB, K., RUDOLPH, G., YAO, X., LUTTON, E., MERELO, J. J. & SCHWEFEL, H.-P. (eds.) *Parallel Problem Solving from Nature PPSN VI: 6th International Conference Paris, France, September 18–20, 2000 Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

DEB, K. & DEB, D. 2014. Analysing mutation schemes for real-parameter genetic algorithms. *International Journal of Artificial Intelligence and Soft Computing,* 4**,** 1-28.

DEB, K. & GOYAL, M. 1996. A combined genetic adaptive search (GeneAS) for engineering design. *Computer Science and informatics,* 26**,** 30-45.

DEB, K., PRATAP, A., AGARWAL, S. & MEYARIVAN, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation,* 6**,** 182-197.

DU, Y., CHANG, C.-I., REN, H., CHANG, C.-C., JENSEN, J. O. & D'AMICO, F. M. 2004. New hyperspectral discrimination measure for spectral characterization. *Optical Engineering,* 43**,** 1777-1786.

ERDFELDER, E., FAUL, F. & BUCHNER, A. 1996. GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers,* 28**,** 1-11.

FYFE, S. Seagrass species: are they spectrally distinct? *Geoscience and Remote Sensing Symposium, 2001. IGARSS'01. IEEE 2001 International,* 2001. IEEE, 2740-2742.

GOLDBERG, D. E. 1988. *Genetic algorithms in search, optimization, and machine learning,* Reading, Mass. : Addison-Wesley Pub. Co., 1989.

GOLDBERG, D. E. 1991. Real-coded genetic algorithms, virtual alphabets, and blocking. *Complex systems,* 5**,** 139-167.

GUALTIERI, J. A. & CROMP, R. F. Support vector machines for hyperspectral remote sensing classification. *27th AIPR Workshop: Advances in Computer-Assisted Recognition,* 1999. International Society for Optics and Photonics, 221-233.

HAIDER, N. 2011. *Identification of plant species using traditional and molecular-based methods*.

HERRERA, F., LOZANO, M. & VERDEGAY, J. L. 1997. Fuzzy connectives based crossover operators to model genetic algorithms population diversity. *Fuzzy Sets and Systems,* 92**,** 21-30.

HEWITT, C. L., CAMPBELL, M., THRESHER, R. & MARTIN, R. 1999. *Marine biological invasions of Port Phillip Bay, Victoria,* CSIRO Marine Research Hobart, Australia.

HIREMATH, P. & PUJARI, J. Content based image retrieval using color, texture and shape features. *Advanced Computing and Communications, 2007. ADCOM 2007. International Conference on,* 2007. IEEE, 780-784.

HOLLAND, J. H. 1975. *Adaptation in Natural and Artificial Systems,* An Arbor, University of Michigan Press.

HORN, J., NAFPLIOTIS, N. & GOLDBERG, D. E. A niched Pareto genetic algorithm for multiobjective optimization. *Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence,* 27-29 Jun 1994 1994. 82-87 vol.1.

HUISMAN, J., 2011, Halodue univervis, AlgaeBase. Available: http://img.algaebase.org/images/AC100CF10ccbb25EAESvQ17B86C2/r7wVcNfwiD7t.jpg [Accessed 1 Feb 2018].

KESHAVA, N. 2004. Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries. *IEEE Transactions on Geoscience and remote sensing,* 42**,** 1552-1565.

KONAK, A., COIT, D. W. & SMITH, A. E. 2006. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering & System Safety,* 91**,** 992-1007.

KUTSER, T. & JUPP, D. L. B. 2006. On the possibility of mapping living corals to the species level based on their optical signatures. *Estuarine, Coastal and Shelf Science,* 69**,** 607-614.

KUTSER, T. & METSAMAA, L. 2006. Spectral library of macroalgae and benthic substrates in Estonian coastal waters. *Proc. Estonian Acad. Sci. Biol. Ecol,* 55**,** 329-340.

MA, D., LIU, J., HUANG, J., LI, H., LIU, P., CHEN, H. & QIAN, J. 2016. Spectral Similarity Assessment Based on a Spectrum Reflectance-Absorption Index and Simplified Curve Patterns for Hyperspectral Remote Sensing. *Sensors (Basel, Switzerland),* 16**,** 152.

MCKENZIE, L., 2007, Cymadocea serrulata, Digital Image, Queensland Department of Primary Industries & Fisheries. Available: https://researchonline.jcu.edu.au/8566/1/8566_Waycott_et_al_2007.pdf [Accessed 1 Feb 2018].

NOMURA, T. & MIYOSHI, T. 1996. Numerical coding and unfair average crossover in GA for fuzzy rule extraction in dynamic environments. *Fuzzy Logic, Neural Networks, and Evolutionary Computation***,** 55-72.

PARINGIT, E. C., NADAOKA, K., FORTES, M. D., HARII, S., TAMURA, H., MITSUI, J. & STRACHAN, J. J. Multiangular and hyperspectral reflectance modeling of seagrass beds for remote sensing studies. *Geoscience and Remote Sensing Symposium, 2003. IGARSS'03. Proceedings. 2003 IEEE International,* 2003. IEEE, 2128-2130.

PEARSON, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* 2**,** 559-572.

POCKLINGTON, J., 2011, Seagrass Zostera Muelleri, Taxonomic Toolkit for marine life of Port Phillip Bay, Museum Victoria. Available: http://portphillipmarinelife.net.au/species/12323 [Accessed 9 May 2017].

RICHARDSON, J. T., PALMER, M. R., LIEPINS, G. E. & HILLIARD, M. Some guidelines for genetic algorithms with penalty functions. *Proceedings of the third international conference on Genetic algorithms*, 1989. Morgan Kaufmann Publishers Inc., 191-197.

ROELFSEMA, C. M., PHINN, S. R. & JOYCE, K. 2016. Spectral reflectance library of algal, seagrass and substrate types in Moreton Bay, Australia. PANGAEA.

ROJAS, I., GONZALEZ, J., POMARES, H., MERELO, J. J., CASTILLO, P. A. & ROMERO, G. 2002. Statistical analysis of the main parameters involved in the design of a genetic algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 32**,** 31-37.

ROSENBERG, R., BLOMQVIST, M., C NILSSON, H., CEDERWALL, H. & DIMMING, A. 2004. Marine quality assessment by use of benthic species-abundance distributions: a proposed new protocol within the European Union Water Framework Directive. *Marine Pollution Bulletin*, 49**,** 728-739.

SCHAFFER, J. D. Multiple objective optimization with vector evaluated genetic algorithms. *Proceeding of the First International Conference on Genetic Algorithms and Their Applications, 1985*, 1985. LawrenceErlbaumAssociates, Inc., Publishers.

SCHWARZ, J. & STAENZ, K. 2001. Adaptive threshold for spectral matching of hyperspectral data. *Canadian Journal of Remote Sensing*, 27**,** 216-224.

SESHADRI, A. 2006. *Multi-objective optimization using evolutionary algorithms (MOEA)* [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/10429 [Accessed 1 Feb 2018].

SHAHRIARI, H., RANJBAR, H., HONARMAND, M. & CARRANZA, E. J. M. 2014. Selection of Less Biased Threshold Angles for SAM Classification Using the Real Value–Area Fractal Technique. *Resource Geology*, 64**,** 301-315.

SRINIVAS, N. & DEB, K. 1994. Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation*, 2**,** 221-248.

VAN DER MEER, F. 2006. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. *International journal of applied earth observation and geoinformation*, 8**,** 3-17.

VAN DER MEER, F. & BAKKER, W. 1997. CCSM: Cross correlogram spectral matching. *International Journal of Remote Sensing,* 18**,** 1197-1201.

VAPNIK, V. N. 1995. *The Nature of Statistical Learning Theory,* Springer science & business media.

VOIGT, H.-M., MÜHLENBEIN, H. & CVETKOVIC, D. Fuzzy recombination for the breeder genetic algorithm. *Proc. Sixth Int. Conf. on Genetic Algorithms,* 1995.

VOLENT, Z., JOHNSEN, G. & SIGERNES, F. 2009. Microscopic hyperspectral imaging used as a bio-optical taxonomic tool for micro-and macroalgae. *Applied optics,* 48**,** 4170-4176.

ZITZLER, E. 1999. *Evolutionary algorithms for multiobjective optimization: Methods and applications.* Doctoral Thesis, Swiss Federal Institute of Technology Zurich.